

Electronic Patent Application Fee Transmittal				
Application Number:		17542193		
Filing Date:				
Title of Invention:		Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
First Named Inventor/Applicant Name:		Gyeongin Yu		
Filer:		Jae Yeon Baek/Carol Bergeron		
Attorney Docket Number:		39192-50678/US		
Filed as Small Entity				
Filing Fees for Utility under 35 USC 111(a)				
Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
LATE FILING FEE FOR OATH OR DECLARATION	2051	1	80	80
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Extension-of-Time:				
Miscellaneous:				
Total in USD (\$)				80

Electronic Acknowledgement Receipt

EFS ID:	44434450
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Yeon Baek
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	03-DEC-2021
Filing Date:	
Time Stamp:	18:39:05
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	\$80
RAM confirmation Number	E2021B3I39403768
Deposit Account	192555
Authorized User	Jae Yeon Baek

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)
 37 CFR 1.20 (Post Issuance fees)
 37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Fee Worksheet (SB06)	fee-info.pdf	38022	no	2
			6c1d7045b8058937fd51b9b6cb0ef552fa42d0e		

Warnings:

Information:

Total Files Size (in bytes):	38022
------------------------------	-------

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

**CERTIFICATION AND REQUEST FOR PRIORITIZED EXAMINATION
UNDER 37 CFR 1.102(e) (Page 1 of 1)**

First Named Inventor:	Gyeongin Yu	Nonprovisional Application Number (if known):	39192-50678/US
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS PRIORITIZED EXAMINATION FOR THE ABOVE-IDENTIFIED APPLICATION.

1. The processing fee set forth in 37 CFR 1.17(i)(1) and the prioritized examination fee set forth in 37 CFR 1.17(c) have been filed with the request. The publication fee requirement is met because that fee, set forth in 37 CFR 1.18(d), is currently \$0. The basic filing fee, search fee, and examination fee are filed with the request or have been already been paid. I understand that any required excess claims fees or application size fee must be paid for the application.
2. I understand that the application may not contain, or be amended to contain, more than four independent claims, more than thirty total claims, or any multiple dependent claims, and that any request for an extension of time will cause an outstanding Track I request to be dismissed.
3. The applicable box is checked below:

I. ☒ Original Application (Track One) - Prioritized Examination under § 1.102(e)(1)

- i. (a) The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a). This certification and request is being filed with the utility application via EFS-Web.
---OR---
(b) The application is an original nonprovisional plant application filed under 35 U.S.C. 111(a). This certification and request is being filed with the plant application in paper.
- ii. An executed inventor's oath or declaration under 37 CFR 1.63 or 37 CFR 1.64 for each inventor, or the application data sheet meeting the conditions specified in 37 CFR 1.53(f)(3)(i) is filed with the application.

II. ☐ Request for Continued Examination - Prioritized Examination under § 1.102(e)(2)

- i. A request for continued examination has been filed with, or prior to, this form.
- ii. If the application is a utility application, this certification and request is being filed via EFS-Web.
- iii. The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a), or is a national stage entry under 35 U.S.C. 371.
- iv. This certification and request is being filed prior to the mailing of a first Office action responsive to the request for continued examination.
- v. No prior request for continued examination has been granted prioritized examination status under 37 CFR 1.102(e)(2).

Signature <u>/Jae Yeon Baek/</u>	Date <u>December 3, 2021</u>
Name (Print/Typed) <u>Jae Yeon Baek</u>	Practitioner Registration Number <u>78,258</u>

Note: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required.*

☒ *Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Doc Code: ECOMM.AUTH/ECOMM.WTDW

Doc Description: Internet Communications Authorized/Internet Communications Authorization Withdrawn

PTO/SB/439 (11-15)

**AUTHORIZATION FOR INTERNET
COMMUNICATIONS IN A PATENT
APPLICATION OR REQUEST TO
WITHDRAW AUTHORIZATION FOR
INTERNET COMMUNICATIONS**

Application No.	Not Yet Known
Filing Date	Herewith
First Named Inventor	Gyeongin Yu
Art Unit	Not Yet Known
Examiner Name	Not Yet Known
Practitioner Docket No.	39192-50678/US

To: Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

I. To authorize permission for Internet Communications.

☒ Recognizing that Internet communications are not secure, I hereby authorize the USPTO to communicate with the undersigned and practitioners in accordance with 37 CFR 1.33 and 37 CFR 1.34 concerning any subject matter of this application via video conferencing, instant messaging, or electronic mail. I understand that a copy of these communications will be made of record in the application file. (MPEP 502.03)

II. To withdraw authorization for Internet Communications.

☐ The authorization given on _____, to the USPTO to communicate with the undersigned and any practitioner in accordance with 37 CFR 1.33 and 37 CFR 1.34 concerning any subject matter of this application via Internet communications such as video conferencing, instant messaging, or electronic mail is hereby withdrawn. I understand that the withdrawal is effective when approved rather than when received.

I am the

☐ applicant.

☒ attorney or agent of record. Registration number 78,258.

☐ attorney or agent acting under 37 CFR 1.34. Registration number _____.

/Jae Yeon Baek/
Signature

December 3, 2021
Date

Jae Yeon Baek
Typed or printed name

(650) 335-7142
Telephone Number

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. Juristic entities must be represented by a patent practitioner (see 37 CFR 1.31, which is applicable to any paper filed on or after September 16, 2012, that is presented on behalf of a juristic entity, regardless of application filing date). Submit multiple forms if more than one signature is required, see below*.

☒ * Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76. This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.			

Secrecy Order 37 CFR 5.2:

☐	Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)
--------------------------	---

Inventor Information:

Inventor	1				Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Gyeongin		Yu		
Residence Information (Select One) US Residency ☒ Non US Residency Active US Military Service					
City	Seoul	Country of Residence ⁱ		KR	
Mailing Address of Inventor:					
Address 1		FriendliAI Inc.			
Address 2		Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province			
Postal Code	08826	Country ⁱ	KR		
Inventor	2				Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Geon-Woo		Kim		
Residence Information (Select One) US Residency ☒ Non US Residency Active US Military Service					
City	Seoul	Country of Residence ⁱ		KR	
Mailing Address of Inventor:					
Address 1		FriendliAI Inc.			
Address 2		Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province			
Postal Code	08826	Country ⁱ	KR		
Inventor	3				Remove
Legal Name					

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Prefix	Given Name	Middle Name	Family Name	Suffix
	Joo Seong		Jeong	
Residence Information (Select One) US Residency ☒ Non US Residency Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:

Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	
Inventor	4	Remove		
Legal Name				

Prefix	Given Name	Middle Name	Family Name	Suffix
	Soojeong		Kim	
Residence Information (Select One) US Residency ☒ Non US Residency Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:

Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	
Inventor	5	Remove		
Legal Name				

Prefix	Given Name	Middle Name	Family Name	Suffix
	Byung-Gon		Chun	
Residence Information (Select One) US Residency ☒ Non US Residency Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:

Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button.

Add

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

☐ An Address is being provided for the correspondence information of this application.

Customer Number	00758		
Email Address	ptoc@fenwick.com	Add Email	Remove Email

Application Information:

Title of the Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
Attorney Docket Number	39192-50678/US	Small Entity Status Claimed	☒
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	12	Suggested Figure for Publication (if any)	

Filing By Reference:

Only complete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32). Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
Please Select One:	☒ Customer Number	☐ US Patent Practitioner	☐ Limited Recognition (37 CFR 11.9)
Customer Number	00758		

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, 365(c), or 386(c) or indicate National Stage entry from a PCT application. Providing benefit claim information in the Application Data Sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78.

When referring to the current application, please leave the "Application Number" field blank.

Prior Application Status		Remove
Application Number	Continuity Type	Prior Application Number
Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the Add button.		Add

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55. When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX)ⁱ the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(i)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

Application Number	Country ⁱ	Filing Date (YYYY-MM-DD)	Access Code ^j (if applicable)
Additional Foreign Priority Data may be generated within this form by selecting the Add button.			Add

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

- ☐ This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013.
- NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Authorization or Opt-Out of Authorization to Permit Access:

When this Application Data Sheet is properly signed and filed with the application, applicant has provided written authority to permit a participating foreign intellectual property (IP) office access to the instant application-as-filed (see paragraph A in subsection 1 below) and the European Patent Office (EPO) access to any search results from the instant application (see paragraph B in subsection 1 below).

Should applicant choose not to provide an authorization identified in subsection 1 below, applicant **must opt-out** of the authorization by checking the corresponding box A or B or both in subsection 2 below.

NOTE: This section of the Application Data Sheet is **ONLY** reviewed and processed with the **INITIAL** filing of an application. After the initial filing of an application, an Application Data Sheet cannot be used to provide or rescind authorization for access by a foreign IP office(s). Instead, Form PTO/SB/39 or PTO/SB/69 must be used as appropriate.

1. Authorization to Permit Access by a Foreign Intellectual Property Office(s)

A. Priority Document Exchange (PDX) - Unless box A in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the World Intellectual Property Organization (WIPO), and any other foreign intellectual property office participating with the USPTO in a bilateral or multilateral priority document exchange agreement in which a foreign application claiming priority to the instant patent application is filed, access to: (1) the instant patent application-as-filed and its related bibliographic data, (2) any foreign or domestic application to which priority or benefit is claimed by the instant application and its related bibliographic data, and (3) the date of filing of this Authorization. See 37 CFR 1.14(h)(1).

B. Search Results from U.S. Application to EPO - Unless box B in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the EPO access to the bibliographic data and search results from the instant patent application when a European patent application claiming priority to the instant patent application is filed. See 37 CFR 1.14(h)(2).

The applicant is reminded that the EPO's Rule 141(1) EPC (European Patent Convention) requires applicants to submit a copy of search results from the instant application without delay in a European patent application that claims priority to the instant application.

2. Opt-Out of Authorizations to Permit Access by a Foreign Intellectual Property Office(s)

☐ A. Applicant **DOES NOT** authorize the USPTO to permit a participating foreign IP office access to the instant application-as-filed. If this box is checked, the USPTO will not be providing a participating foreign IP office with any documents and information identified in subsection 1A above.

☐ B. Applicant **DOES NOT** authorize the USPTO to transmit to the EPO any search results from the instant patent application. If this box is checked, the USPTO will not be providing the EPO with search results from the instant application.

NOTE: Once the application has published or is otherwise publicly available, the USPTO may provide access to the application in accordance with 37 CFR 1.14.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.			
Applicant	1		Remove
If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.			
		Clear	
☒ Assignee	Legal Representative under 35 U.S.C. 117		Joint Inventor
Person to whom the inventor is obligated to assign.		Person who shows sufficient proprietary interest	
If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:			
Name of the Deceased or Legally Incapacitated Inventor:			
If the Applicant is an Organization check here. ☒			
Organization Name	FriendliAI Inc.		
Mailing Address Information For Applicant:			
Address 1	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu		
Address 2			
City	Seoul	State/Province	
Country	KR	Postal Code	08826
Phone Number		Fax Number	
Email Address			
Additional Applicant Data may be generated within this form by selecting the Add button.			Add

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Applicant	2 Remove		
If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section. Clear			
☒ Assignee	Legal Representative under 35 U.S.C. 117	Joint Inventor	
Person to whom the inventor is obligated to assign.		Person who shows sufficient proprietary interest	
If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:			
Name of the Deceased or Legally Incapacitated Inventor:			
If the Applicant is an Organization check here. ☒			
Organization Name	Seoul National University R&DB Foundation		
Mailing Address Information For Applicant:			
Address 1	1, Gwanak-ro, Gwanak-gu		
Address 2			
City	Seoul	State/Province	
Country	KR	Postal Code	08826
Phone Number		Fax Number	
Email Address			
Additional Applicant Data may be generated within this form by selecting the Add button. Add			

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Assignee	1
Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication. An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication. Remove	
If the Assignee or Non-Applicant Assignee is an Organization check here. ☐	

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	39192-50678/US
		Application Number	
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Prefix	Given Name	Middle Name	Family Name	Suffix

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1			
Address 2			
City		State/Province	
Country ⁱ		Postal Code	
Phone Number		Fax Number	
Email Address			

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button.

Signature:

NOTE: This Application Data Sheet must be signed in accordance with 37 CFR 1.33(b). However, if this Application Data Sheet is submitted with the INITIAL filing of the application and either box A or B is not checked in subsection 2 of the "Authorization or Opt-Out of Authorization to Permit Access" section, then this form must also be signed in accordance with 37 CFR 1.14(c).

This Application Data Sheet **must** be signed by a patent practitioner if one or more of the applicants is a **juristic entity** (e.g., corporation or association). If the applicant is two or more joint inventors, this form must be signed by a patent practitioner, **all** joint inventors who are the applicant, or one or more joint inventor-applicants who have been given power of attorney (e.g., see USPTO Form PTO/AIA/81) on behalf of **all** joint inventor-applicants.

See 37 CFR 1.4(d) for the manner of making signatures and certifications.

Signature	/Jae Yeon Baek/		Date (YYYY-MM-DD)	2021-12-03
First Name	Jae Yeon	Last Name	Baek	Registration Number
				78258

Additional Signature may be generated within this form by selecting the Add button.

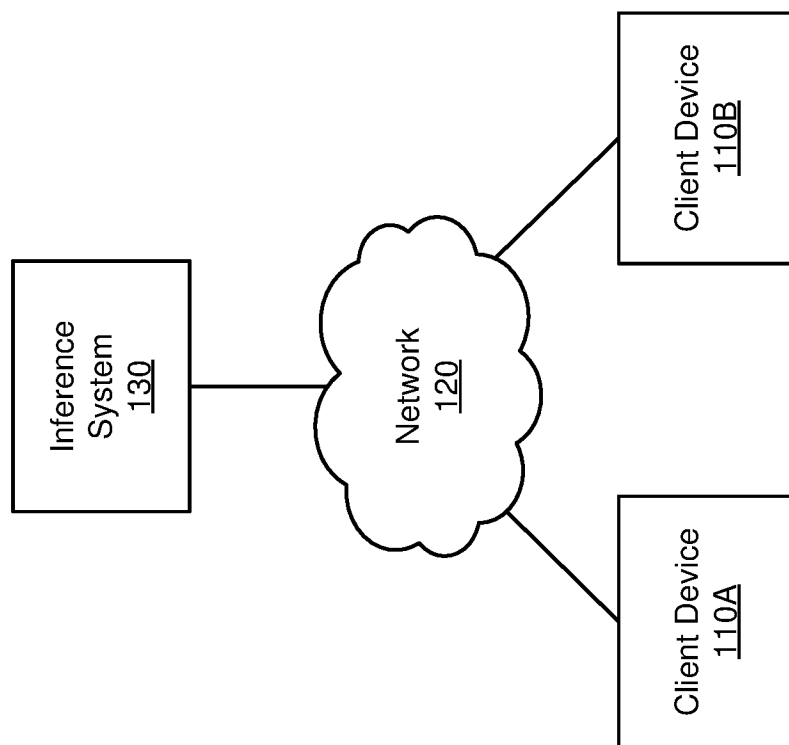
This collection of information is required by 37 CFR 1.76. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 23 minutes to complete, including gathering, preparing, and submitting the completed application data sheet form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.



100

FIG. 1

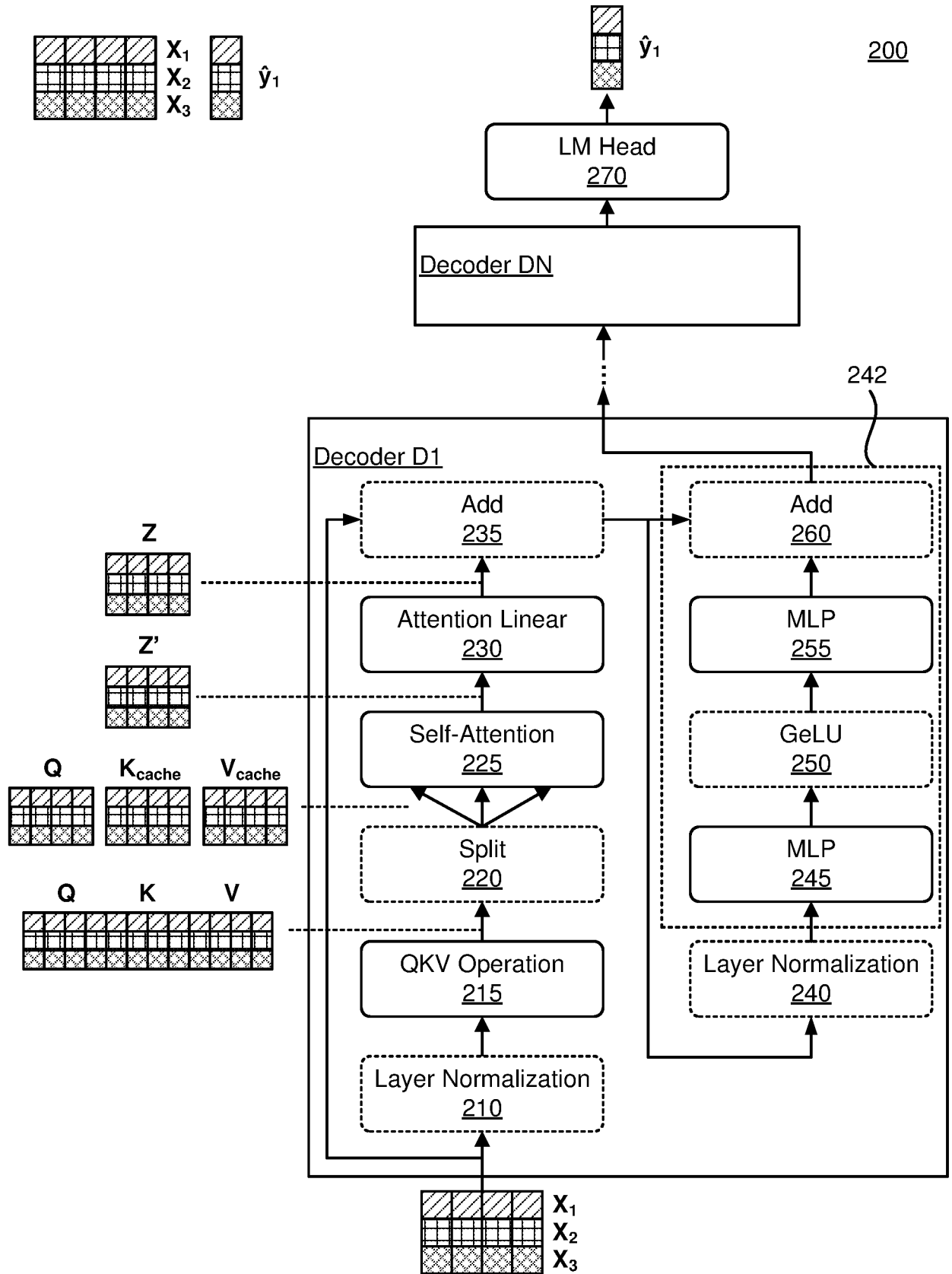


FIG. 2A

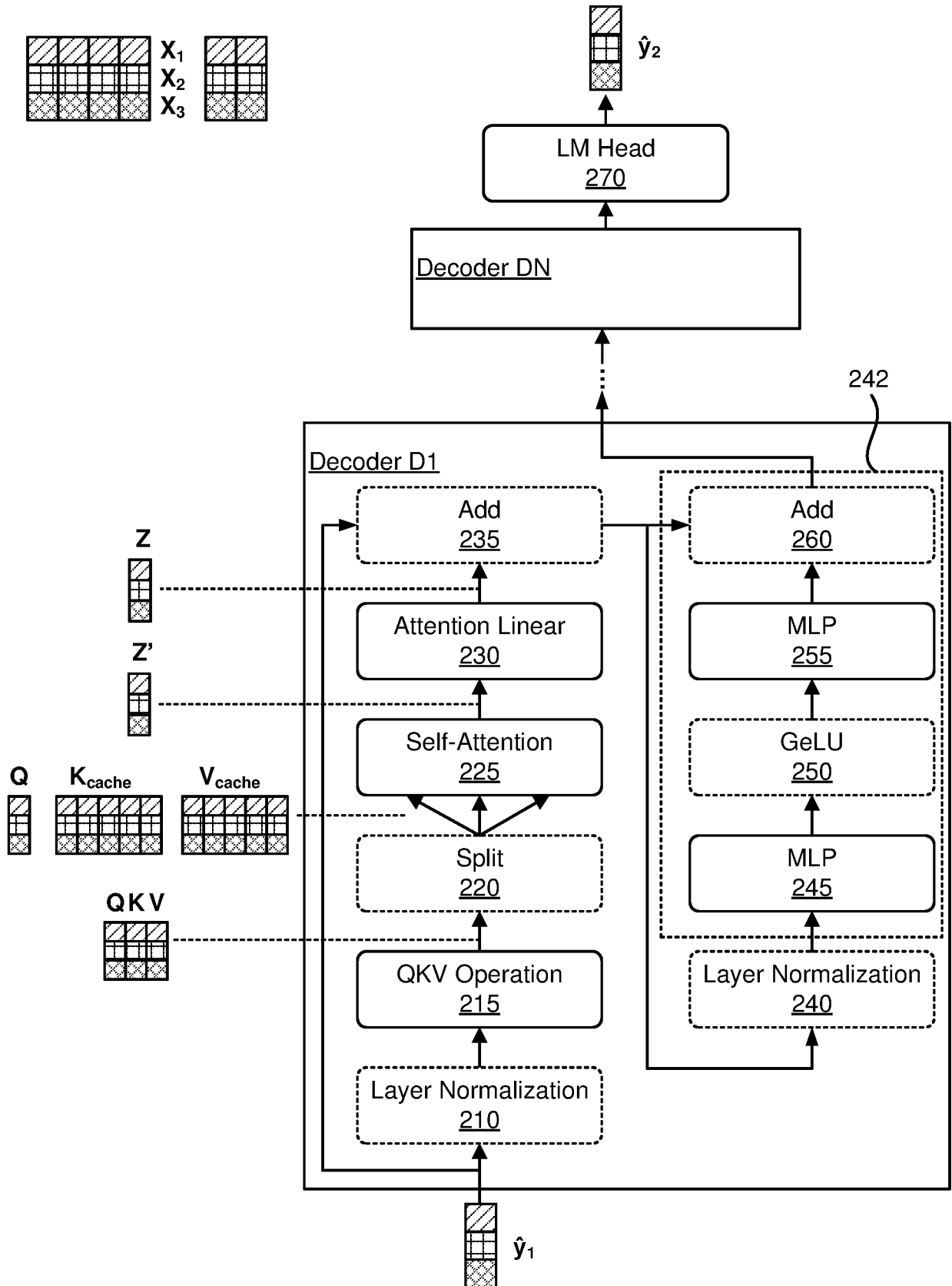


FIG. 2B

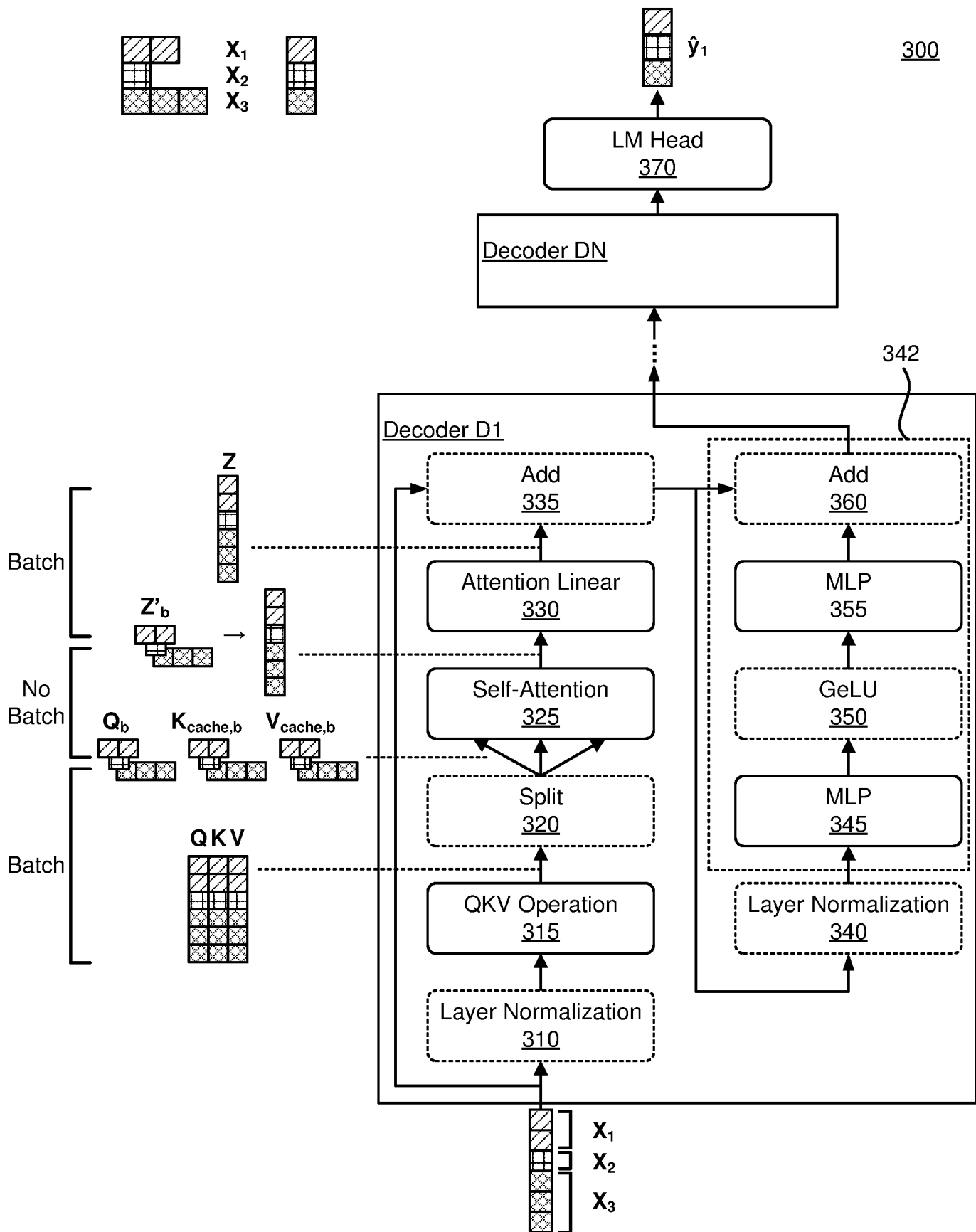


FIG. 3A

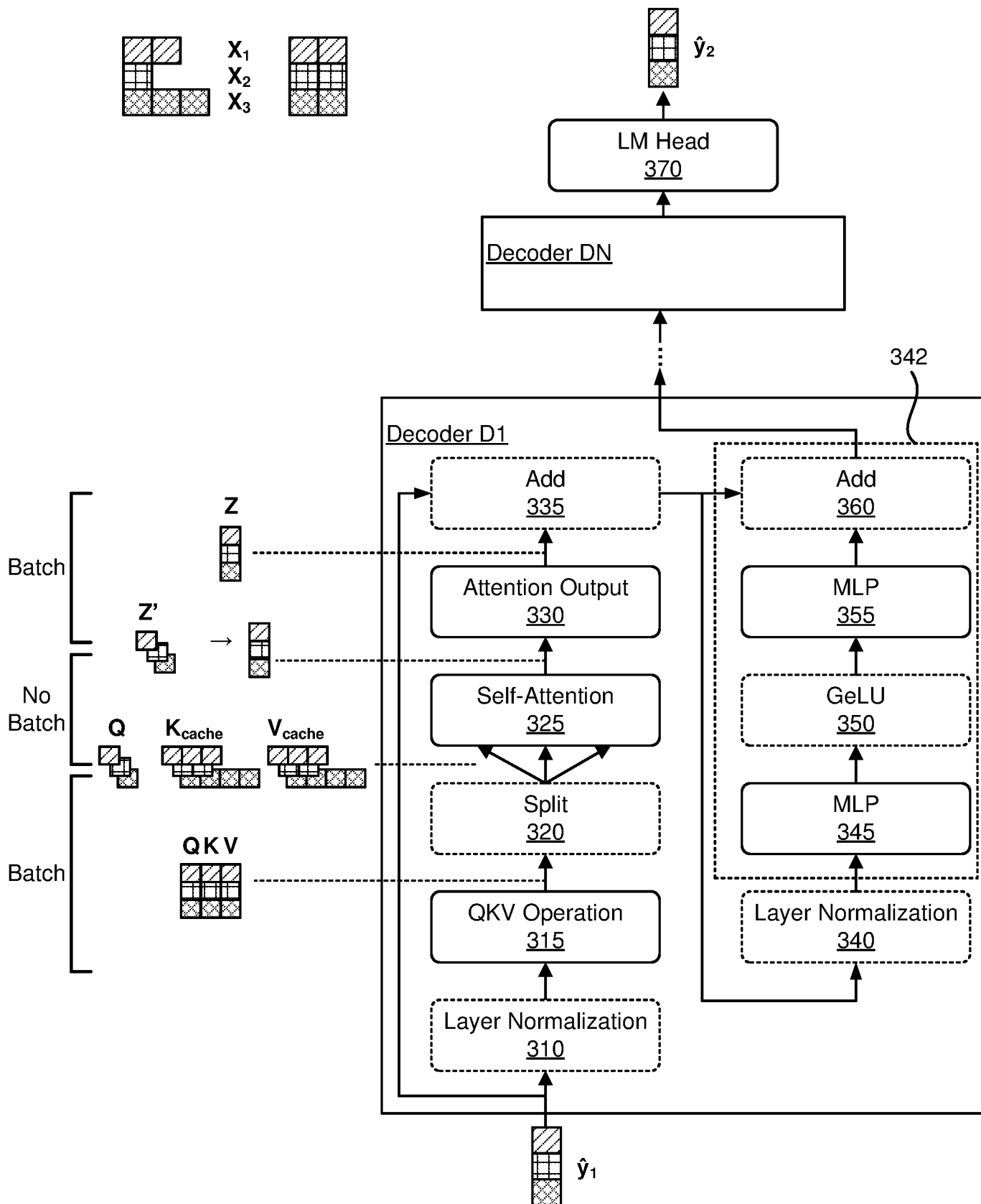


FIG. 3B

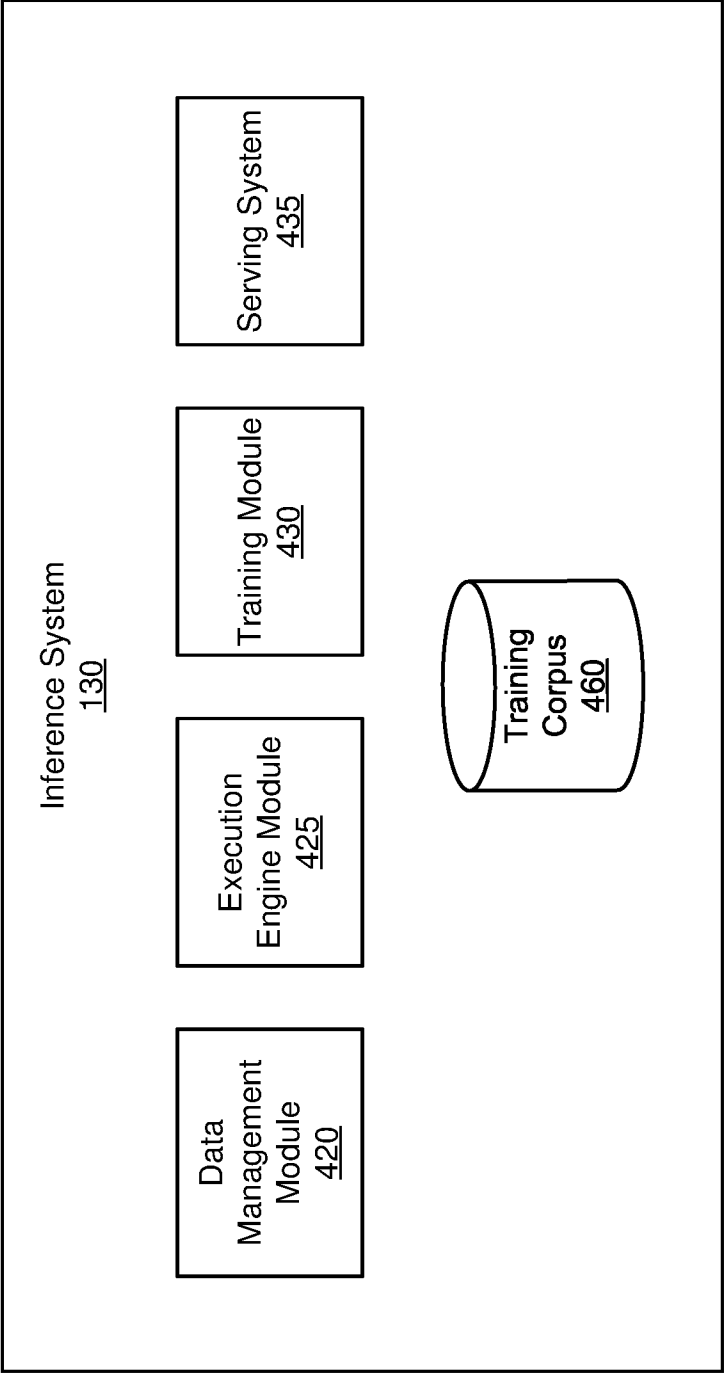


FIG. 4

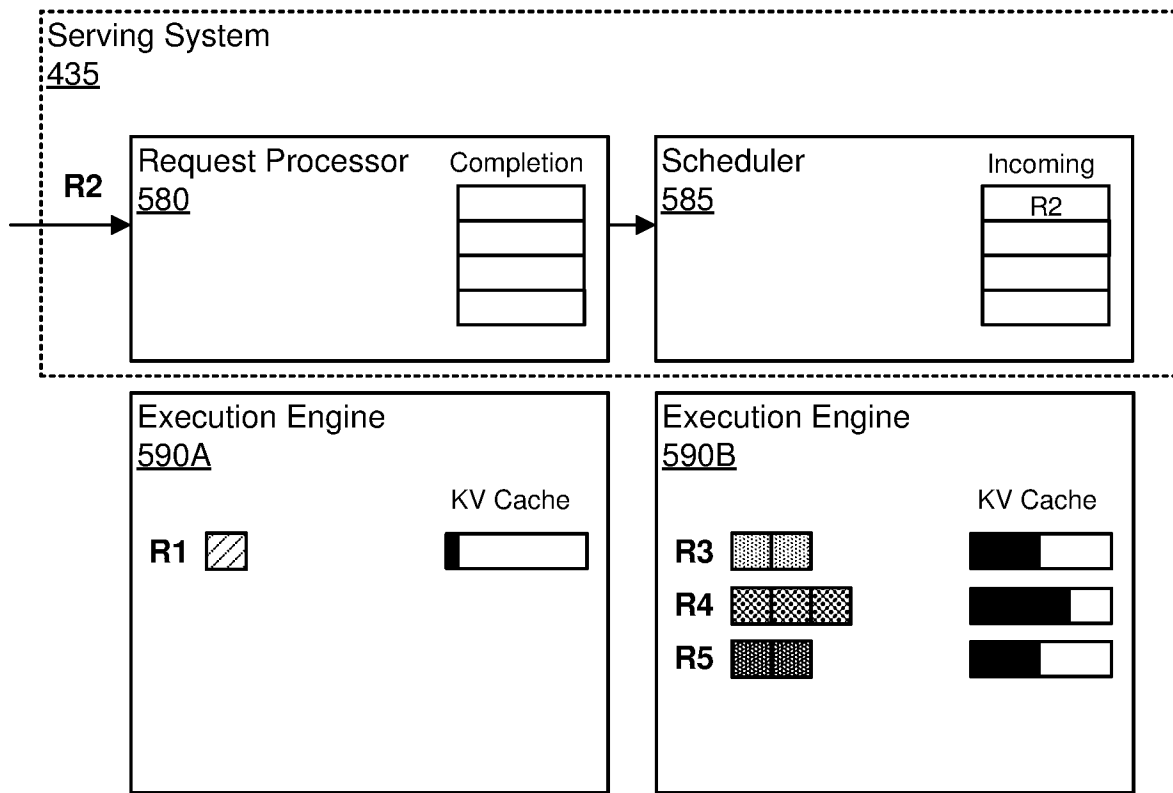


FIG. 5A

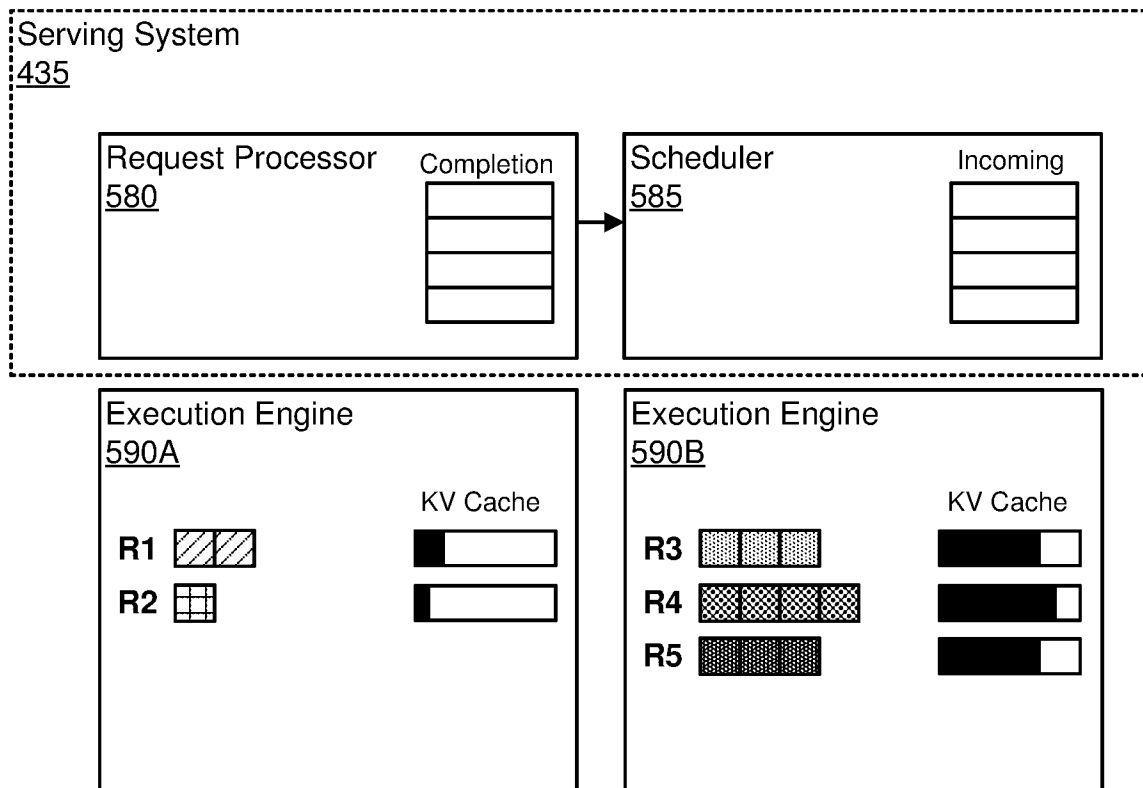


FIG. 5B

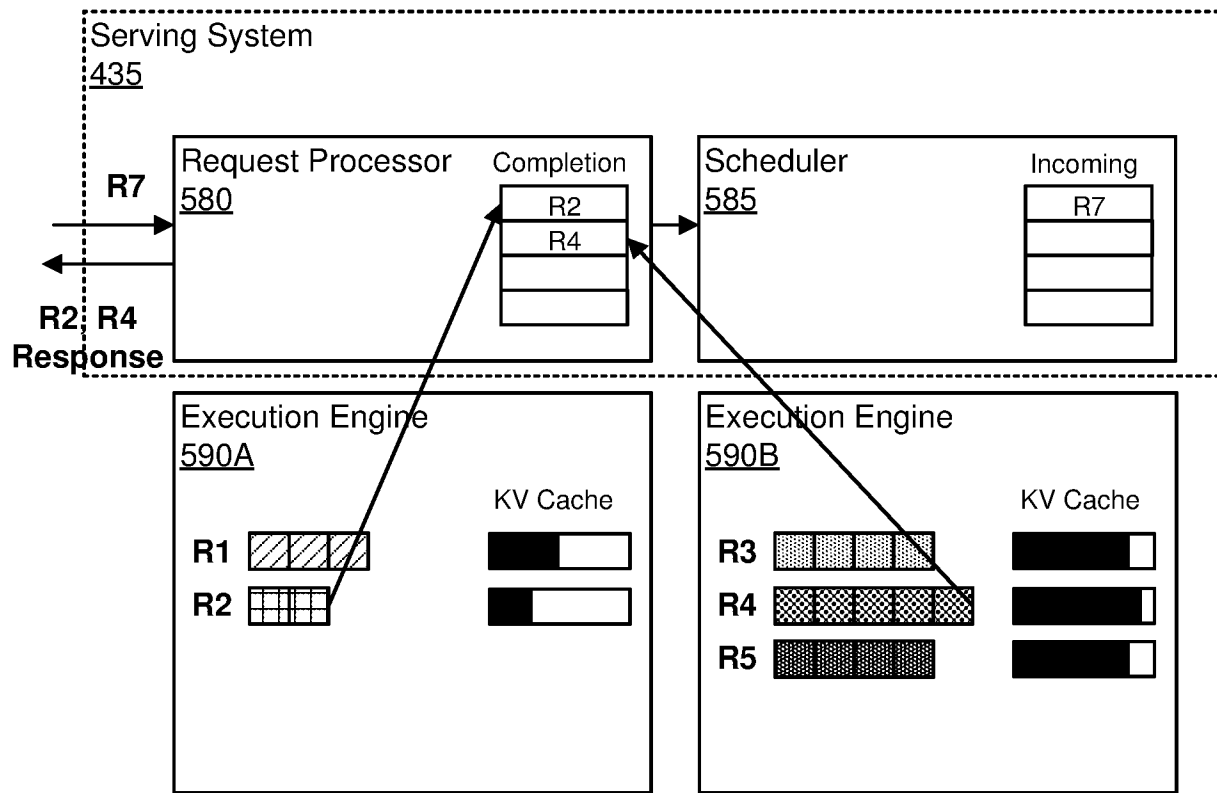


FIG. 5C

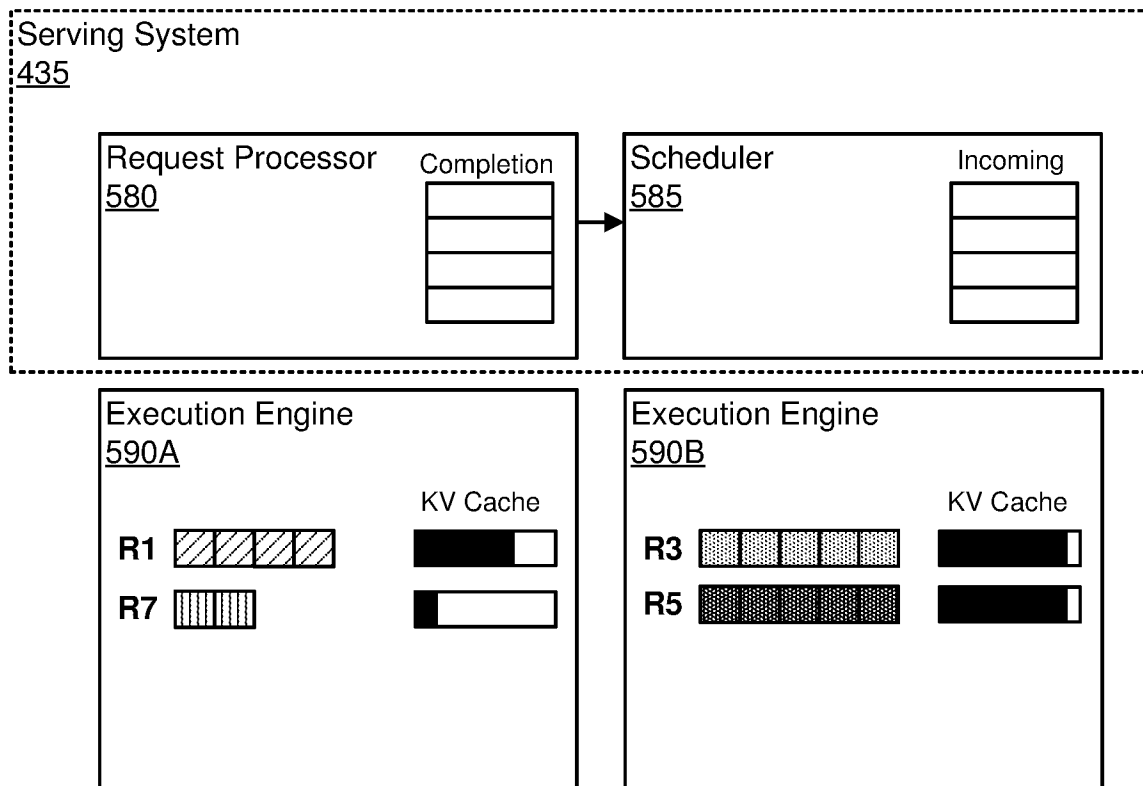


FIG. 5D

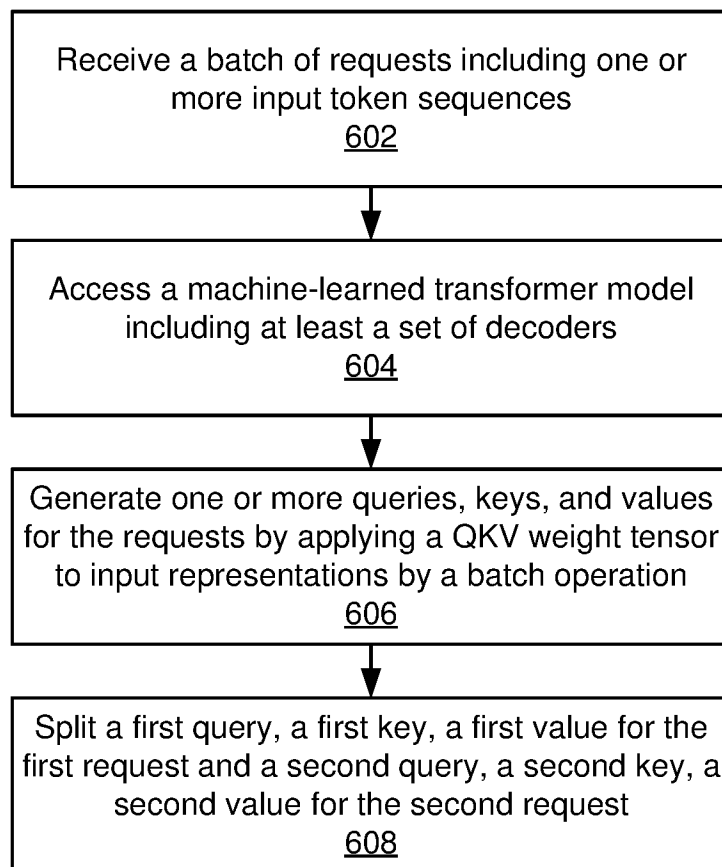


FIG. 6A

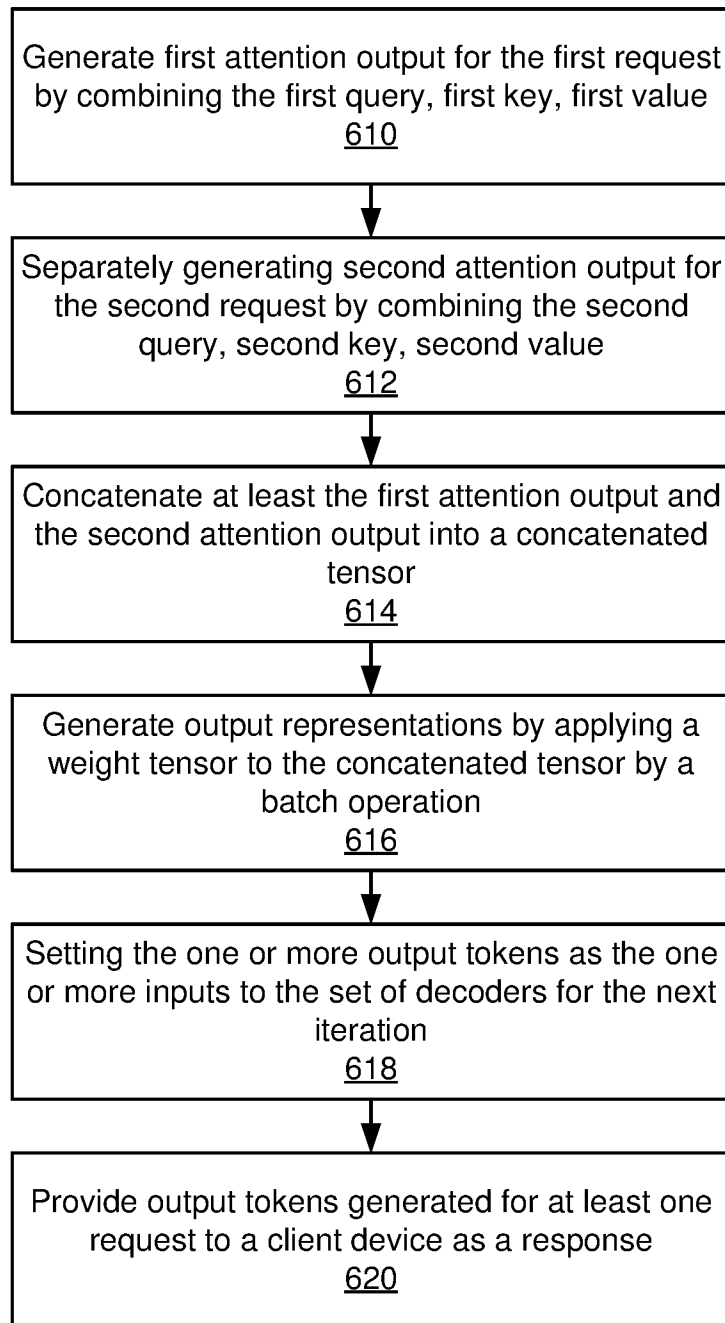


FIG. 6B

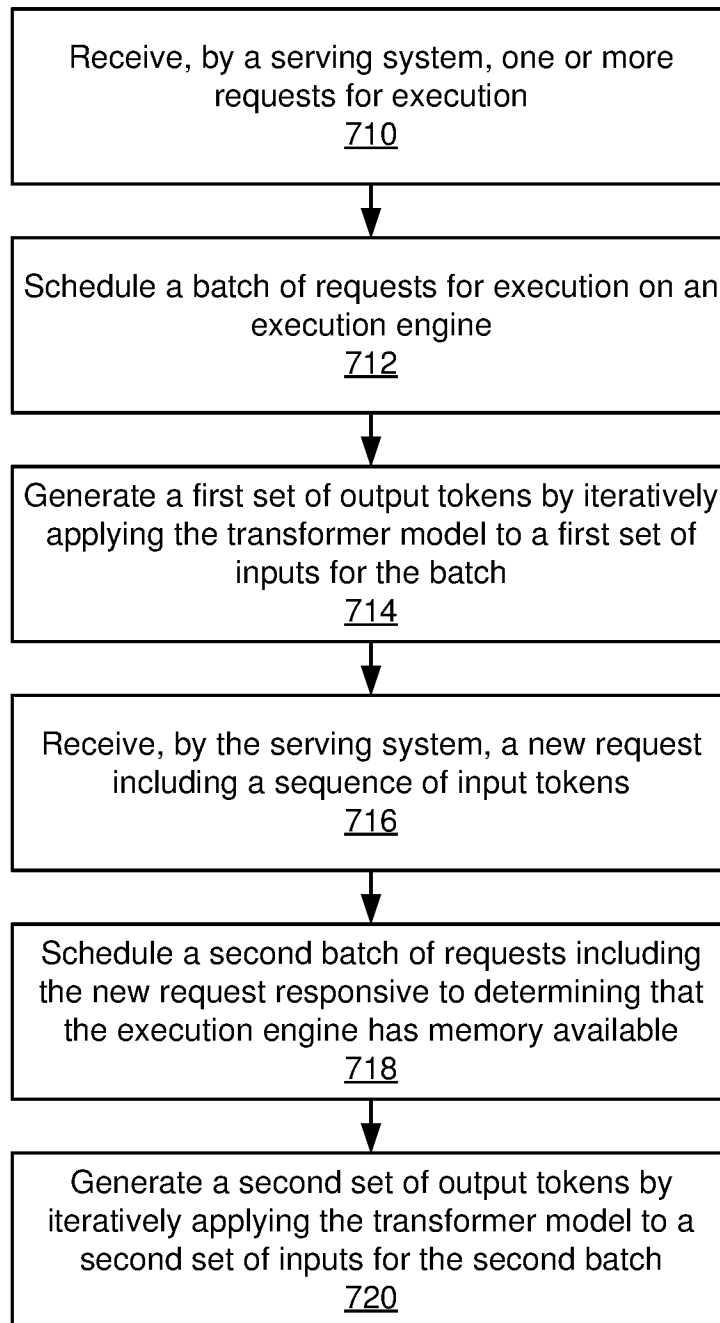


FIG. 7

800

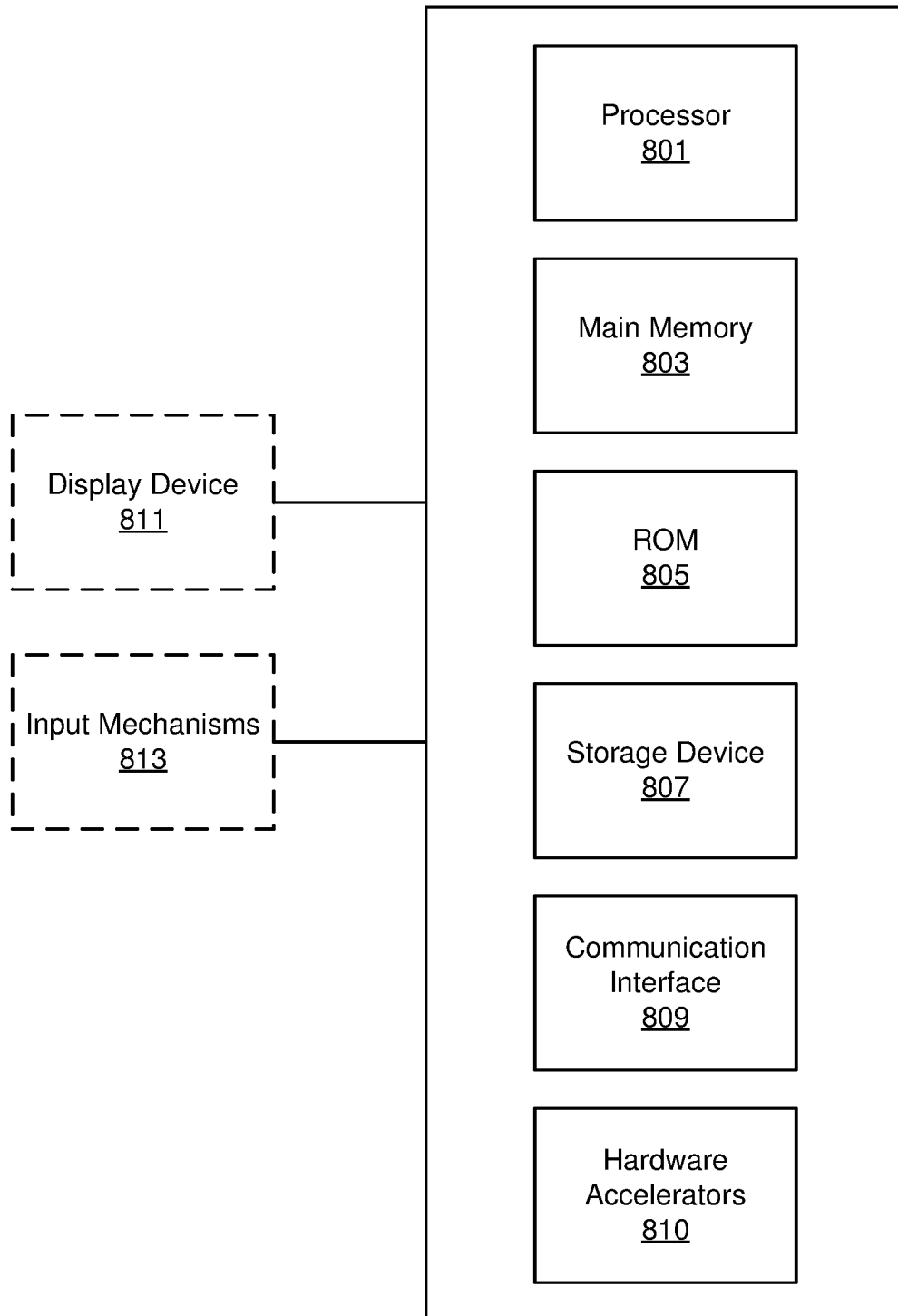


FIG. 8

DYNAMIC BATCHING FOR INFERENCE SYSTEM FOR TRANSFORMER-BASED GENERATION TASKS

Inventors:

Gyeongin Yu
Geon-Woo Kim
Joo Seong Jeong
Soojeong Kim
Byung-Gon Chun

BACKGROUND

[0001] This invention relates generally to machine-learning transformer neural network models, and more particularly to selective batching for transformer models.

[0002] Transformer neural network models are machine-learning models used for a variety of applications, for example, natural language processing (NLP), image processing, or audio processing applications that include sequential data. For example, a transformer model may receive a sequence of input tokens that represent a query and generate a sequence of output tokens that represent a response to the query. As another example, the transformer model may receive a sequence of input tokens that represent a paragraph in German and generate a sequence of output tokens that represents a translation of the paragraph in English. As yet another example, the transformer model may receive a sequence of input tokens that represent a paragraph of text and generate a sequence of output tokens that represent a summarized version of the text.

[0003] Typically, users of client devices submit requests to an inference system. The inference system executes a machine-learning transformer model to inputs (e.g., a sequence of input tokens) of requests to generate outputs (e.g., a sequence of output tokens) for the requests. The inference system may return the outputs to client devices of the requests as a response. In

one instance, the inference system executes the requests on specialized hardware accelerators such as graphics processing units (GPU's) or tensor processing units (TPU's) to improve latency and throughput, especially when the number of parameters of the transformer model is significantly large.

[0004] In one instance, the inference system processes requests in batches to achieve high processor utilization on the accelerators. Specifically, the inference system may process multiple requests in a batch together to exploit the amount of parallel computation units in the hardware accelerators. In many situations, the inputs for requests in a batch are variable in length. For example, the number of input tokens for each request in a batch may be variable in length. However, methods of batching for transformer models often require that the length of data for multiple requests in a batch be the same to be processed. Thus, it may not be feasible to process a batch of requests with variable lengths or workarounds addressing this problem may result in using more resources compared to processing each request individually.

SUMMARY

[0005] An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing

unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length.

[0006] Specifically, in one embodiment, the inference system receives a batch of requests including one or more input token sequences. A length of a first input token sequence for a first request in the batch may be different from a length of a second input token sequence for a second request. The inference system accesses a transformer model including at least a set of decoders coupled to one another. For one or more iterations, the inference system repeatedly performs the steps of generating one or more output tokens for the requests by applying the set of decoders to one or more inputs for the requests.

[0007] For at least one decoder in the set, the inference system generates one or more queries, one or more keys, and one or more values for the requests by applying a QKV weight tensor to one or more input representations. In one instance, the queries, keys, and values are generated by a batch operation. The inference system splits at least a first query for the first request from the one or more queries, a first key for the first request from the one or more keys, and a first value for the first request from the one or more values. The inference system also splits at least a second query for the second request from the one or more queries, a second key for the second request from the one or more keys, and a second value for the second request from the one or more values.

[0008] The inference system generates a first attention output for the first request by at least combining the first query, the first key, and the first value for the first request. The inference system also separately generates a second attention output for the second request by at least combining the second query, the second key, and the second value for the second request. The inference system concatenates at least the first attention output and the second attention output

into a concatenated tensor and generates one or more output representations by applying a weight tensor to the concatenated tensor. In one instance, the one or more output representations are generated by a batch operation. The inference system sets the one or more output tokens as the one or more inputs to the set of decoders for the next iteration and provides output tokens generated for at least one request to a client device as a response to the at least one request.

[0009] In one embodiment, the inference system performs iteration-level dynamic batching for a transformer model that allows the inference system to dynamically modify a batch of requests being executed on an execution engine. Specifically, in existing batching methods for transformer models, it is difficult to modify a batch of requests once it has started to process on an execution engine. This is because certain methods of batching require the length of the inputs or the length of the internal states to be the same across all requests in the batch.

Therefore, unless new incoming requests have the same length of inputs as the batch of requests being executed on the execution engine, it may be difficult for the inference system to modify the batch to, for example, add or remove new requests to the batch.

[0010] By performing selective batching, the inference system can monitor and modify a batch being processed on the execution engine on an iteration-level and update the batch between iterations as requests get completed and new requests are received. Specifically, at one or more iterations, the inference system can modify the batch being executed on the execution engine by adding new incoming requests to the batch or removing completed requests from the batch.

This is because selective batching allows requests with variable lengths to be processed without restraining the one or more inputs or internal states to the transformer model to same lengths.

This allows the inference system to remove requests in the batch that are completed earlier than others so that the response can be provided to the user faster and allows the inference system to

add new requests to a batch of requests if the execution engine is being under-utilized.

[0011] In one embodiment, a serving system of the inference system receives one or more requests for execution. The serving system may include a request processor and a scheduler each coupled to one or more execution engines for executing a machine-learning transformer model including at least a set of decoders. The scheduler schedules a batch of requests including the one or more requests for execution on an execution engine. The execution engine generates a first set of output tokens by iteratively applying the transformer model to a first set of inputs for the batch of requests. In one instance, applying the transformer model includes applying at least one batch operation to one or more input tensors associated with the batch of requests.

[0012] The serving system may receive a new request from a client device that includes a sequence of input tokens. The scheduler schedules a second batch of requests including the one or more requests and the new request for execution on the execution engine responsive to determining that the execution engine has memory available to execute the second batch of requests. The execution engine generates a second set of output tokens by iteratively applying the transformer model to a second set of inputs for the second batch of requests. The second set of inputs may include the sequence of input tokens for the new request.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] FIG. 1 is a high-level block diagram of a system environment for an inference system, in accordance with an embodiment.

[0014] FIGs. 2A-2B illustrate a method of batching using a machine-learning transformer model, in accordance with an embodiment.

[0015] FIGs. 3A-3B illustrate a method of selective batching using a machine-learning transformer model, in accordance with an embodiment.

[0016] FIG. 4 is a block diagram of an inference system, in accordance with an embodiment.

[0017] FIGs. 5A-5D illustrate a method of dynamic batching for processing requests using a machine-learning transformer model, in accordance with an embodiment.

[0018] FIGs. 6A-6B is a flowchart illustrating a method of selective batching using the transformer model, in accordance with an embodiment.

[0019] FIG. 7 is a flowchart illustrating a method of dynamic batching for processing requests using the transformer model, in accordance with an embodiment.

[0020] FIG. 8 is a diagram illustrating a computer system upon which embodiments described herein may be implemented within the inference system, in accordance with an embodiment.

[0021] The figures depict various embodiments of the present invention for purposes of illustration only. One skilled in the art will readily recognize from the following discussion that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the invention described herein.

DETAILED DESCRIPTION

OVERVIEW

[0022] FIG. 1 is a high-level block diagram of a system environment 100 for an inference system 130, in accordance with an embodiment. The system environment 100 shown by FIG. 1 includes one or more client devices 110A, 110B, a network 120, and an inference system 130.

In alternative configurations, different or additional components may be included in the system

environment 100.

[0023] The inference system 130 receives requests from client devices 110A, 110B to perform tasks using machine-learning models. In one embodiment, the machine-learning models are transformer neural network models. The tasks may include, but are not limited to, natural language processing (NLP), image processing, audio processing applications. Specifically, the transformer model may be appropriate for processing sequential data that can be tokenized into a sequence of input tokens for the request and a sequence of output tokens for the desired response. The inference system 130 receives a request including input data (e.g., text data, image or video data, audio data) and encodes the input data to a set of input tokens. The inference system 130 repeatedly applies the machine-learning transformer model for one or more iterations to generate a set of output tokens. The inference system 130 decodes the set of output tokens to output data and returns the output data as the response to the request. While for applications such as NLP applications, a sequence of input tokens or output tokens are arranged along one dimension (1-D) to represent, for example, a sequence of words, it is appreciated that in other embodiments, a sequence of input tokens or output tokens may be a multi-dimensional sequence. For example, for two-dimensional image data, the sequence of tokens may be a two dimensional (2-D) sequence arranged along both a first direction (e.g., X-axis) and a second direction (e.g., Y-axis), where each token corresponds to a block of one or more pixels in the image.

[0024] In particular, NLP tasks involve using artificial intelligence and machine learning techniques to analyze language and may include a variety of tasks including translation, sentiment analysis, text summarization, auto-correction, and the like. When processing NLP tasks, the inference system 130 receives a request including input text of a sequence of words

(e.g., query) and encodes the input text to a sequence of input tokens that each represent a respective word in a latent space. The inference system 130 repeatedly applies a transformer model for one or more iterations to generate a sequence of output tokens (e.g., response to query). The output tokens are converted to output text as a response to the request.

[0025] For example, a transformer model may receive a sequence of input tokens that represent a query and generate a sequence of output tokens that represent a response to the query. As another example, the transformer model may receive a sequence of input tokens that represent a paragraph in French and generate a sequence of output tokens that represents a translation of the paragraph or sentence in English. As yet another example, the transformer model may receive a sequence of input tokens that represent a paragraph of text and generate a sequence of output tokens that represents a summarized version of the text.

[0026] In one embodiment, the inference system 130 includes one or more execution engines that are built on specialized hardware accelerators such as graphics processing units (GPU's) or tensor processing units (TPU's). The requests are executed on the execution engines. Specifically, execution of machine-learning neural network models, such as transformer models, involve a significant number of operations, such as tensor multiplication between input data and high-dimensional weight tensors that can be computationally intensive. The hardware accelerators of the execution engines may be optimized to perform these operations efficiently by parallel processing, leading to significant improvement in latency or throughput when the number of parameters in the transformer model are large.

[0027] The hardware of the inference system 130 may include one or more central processing unit (CPU) cores, CPU memory (e.g., DRAM), data storage, one or more execution engines (e.g., GPU devices). Each execution engine may include a set of cores (e.g., GPU

cores) coupled to local memory (e.g., GPU memory), and may be composed of one or more hardware accelerators. In addition, the inference system 130 may be composed of multiple hardware components and components for configuring a network to connect the various components across the multiple hardware components together such that the components can coordinate with each other to process requests. For example, one execution engine may communicate with multiple hardware accelerators on multiple machines. An execution engine may process data that is stored on its local memory. Specifically, during training or inference of the transformer model, data required for inference or training is read from an input file in the data storage by the CPU or across the network 120 from, for example, a client device 110, moved to local memory of an execution engine, and processed by the execution engine. The results of the processing are retrieved by the CPU.

[0028] In one embodiment, the inference system 130 processes requests by batches to achieve higher processor utilization on the hardware accelerators. Specifically, the inference system 130 processes multiple requests in a batch together to exploit the amount of parallel computation units in the execution engines. In such an embodiment, the inference system 130 receives multiple requests each associated with an input token sequence. The inference system 130 iteratively applies the transformer model to the batch of requests to generate the output tokens for the requests together. In one instance, batching for a transformer model is made possible by grouping requests that have the same length of input token sequences together or at each iteration, treating requests in the batch as if they all had the same input token sequence lengths as the request with the shortest length.

Transformer Model with Batching

[0029] FIGs. 2A-2B illustrate a method of batching using a machine-learning transformer

model 200, in accordance with an embodiment. In particular, the transformer model 200 is associated with a set of parameters determined through a training process. The transformer model 200 is coupled to receive one or more inputs (e.g., input token sequences or output tokens of previous iteration) and generate one or more outputs (e.g., output token predictions). Specifically, in one embodiment, a request is processed by applying the transformer model to the input token sequence of the request in one iteration of an “encoding phase,” and applying the transformer model to an output token sequence generated at a previous iteration for one or more iterations of a “decoding phase.” In a decoding phase, the output tokens of a previous iteration are set as inputs for a next iteration, and the process is repeated for additional iterations of the decoding phase until an end token (e.g., period “.” or an “<end>” indicator) is generated for each request in the batch signaling that all predictions have been generated for a respective request. In other embodiments, the termination criteria for a request may be to terminate after a maximum number of output tokens are generated or may be other types of conditions that are specified by, for example, the inference system 130 or users of the client devices 110.

[0030] FIG. 2A illustrates an encoding phase for the transformer model 200, in which the set of input token sequences are processed to generate one or more output tokens. In the example shown in FIG. 2A, the inference system 130 processing requests for a chatbot receives a first request as the question “what is your name?,” a second request as the question “what is the time?,” and a third request as the question “how do I pay?” The inference system 130 encodes each of the requests as a respective set of input token sequences. The first request is encoded to an input token sequence $\mathbf{X}_1$, the second request is encoded to an input token sequence $\mathbf{X}_2$, and the third request is encoded to an input token sequence $\mathbf{X}_3$, each request being illustrated with a different fill pattern in the figures. Each input token sequence in FIG. 2A is a one-dimensional

sequence in which a sequence of tokens are arranged along a single dimension (e.g., X-direction). However, as described above with respect to FIG. 1, it is appreciated that in other embodiments, a sequence of tokens may be arranged as a multi-dimensional sequence.

[0031] As shown in FIG. 2A, since each request includes four words, each input token sequence includes four tokens each token representing a respective word. For example, input token sequence X_1 for the first request is represented by four squares that represent words “what,” “is,” “your,” “name.” Specifically, while each word is mapped to a single square, in practice, the inference system 130 represents a token for a word as an embedding that represents the word in a multi-dimensional latent space. Thus, while each input token sequence is visually illustrated as a two-dimensional 1×4 tensor in FIG. 2A, in practice, each input token sequence may be represented as a three-dimensional tensor $1 \times 4 \times H$ where H is the dimension of an embedding (e.g., direction going in or out of the page). Moreover, while each token (input token or output token) is mapped to one word for the remainder of the specification, this is merely an example, and it is appreciated that in other embodiments, each token may be mapped to different text units, combination of text units, and the like. For example, in other embodiments, each token may be mapped to a text unit of multiple words, paragraphs, sentences, n-grams or may be mapped to a punctuation mark (e.g., “?”, “!,” “:”) in addition to text units.

[0032] In one embodiment, the transformer model 200 includes a set of N decoders $D_1, D_2, \dots, D_N$. A decoder is coupled to receive a set of input representations and generate a set of output representations. For example, the first decoder D_1 is coupled to receive the one or more inputs to the transformer model 200 as the set of input representations and generate a set of output representations. Each subsequent decoder is coupled to receive the set of output representations of a previous decoder and generate another set of output representations. For

example, the second decoder D2 placed after the first decoder D1 is coupled to receive the set of output representations generated by the first decoder D1 and generate another set of output representations. This process is repeated until the set of output representations for the final decoder are generated.

[0033] The transformer model 200 may also include a LM head block 270 that is coupled to receive the set of output representations from the final decoder DN and generate one or more output tokens as the outputs for the current iteration. In the example illustrated in FIG. 2A, the LM head 270 receives the set of output representations from a final decoder DN and generates one or more output tokens for the batch of requests in the forms of a tensor $\hat{\mathbf{y}}_1$. Specifically, tensor $\hat{\mathbf{y}}_1$ includes a first output token for the first request that represents the first word prediction “my” for the response, a second output token for the second request that represents the word prediction “the” for the response, and a third output token for the third request that represents the word prediction “you” for the response.

[0034] FIG. 2B illustrates a decoding phase for the transformer model 200, in which the previous output tokens are processed as inputs to generate one or more additional output tokens. Different from the encoding phase, the transformer model 200 is applied to output tokens generated at a previous iteration as inputs to generate the additional output tokens. As shown in FIG. 2B, the transformer model 200 at the next iteration receives the output tokens $\hat{\mathbf{y}}_1$ generated at the previous iteration and generates additional output tokens $\hat{\mathbf{y}}_2$ for the requests as outputs. The decoding phase is repeated until the termination criteria is satisfied for each request in the batch.

[0035] The transformer model 200, and in particular, each decoder D_i in the set of decoders D1, D2, . . . , DN include one or more blocks that each represent a respective operation, in

particular a tensor operation. Specifically, a block in the transformer model 200 may contain data for performing an operation to one or more input tensors for a request to generate one or more output tensors for the request. When an operation is executed on an execution engine, the one or more input tensors and associated data (e.g., weight tensors) for the operation may be read from the local memory of the execution engine. The operation may be executed using the one or more input tensors and the associated data to generate one or more output tensors. The output tensors are provided to the CPU, to another execution engine, or may be stored on the local memory of the execution engine such that it can be used for the next operation.

[0036] In one embodiment, the operations of the transformer model 200 are configured as batch operations in which data for a batch of requests are processed together. A batch operation is coupled to receive one or more input tensors that are each concatenations of input data for multiple requests in a batch. The batch operation generates one or more output tensors that are each concatenations of output data for the multiple requests in the batch by applying the respective operation to the one or more input tensors. For a batch operation, an input tensor may be a concatenation of the input tensor for each respective request in the batch. Thus, the input tensor for the batch operation is a single, larger tensor that coalesces the input tensors for each request in the batch across an additional batch dimension.

[0037] Similarly, an output tensor from a batch operation may be represented as the concatenation of output tensors for each respective request in the batch. Thus, while an output tensor for an individual request can also be generated by applying the operation to the input tensor for the request, an output tensor generated from a batch operation using a concatenated input tensor is also a single, larger tensor that coalesces the output tensors for each request in the batch across the batch dimension.

[0038] Moreover, certain operations, such as tensor multiplication operations, involve multiplying a set of parameters (e.g., weight tensors) of the transformer model 200 with the input tensors for the batch operation. In some instances, when the requests are processed individually, a weight tensor is read from the local memory of a hardware accelerator each time an input tensor for a request is multiplied by the weight tensor. In contrast, when executing the tensor multiplication as a batch operation, the set of parameters are read once and re-used for operation on the concatenated input tensor. This can lead to significant improvement in throughput compared to processing the requests individually. However, in the batching process of FIGs. 2A-2B, each request may be required to have the same length in an input tensor and an internal state tensor since the input data for the batch are processed as a single tensor.

[0039] As shown in FIG. 2A, a decoder in the transformer model 200 includes a first layer normalization block 210, a QKV operation block 215, a split block 220, a self-attention block 225, an attention linear block 230, a first add block 235, a second layer normalization block 240, a first MLP block 245, a GeLU block 250, a second MLP block 255, and a second add block 260. In one embodiment, each of the blocks in the decoder are performed on an execution engine as a batch operation, in which input tensors for the batch operation are a concatenation of the input data for the batch of requests and the output tensors are a concatenation of the output data for the batch of requests. While the operations in the first decoder D1 are described as an example, it is appreciated that the remaining decoders in the set may also include similar operations as the first decoder D1.

[0040] Specifically, the inference system 130 concatenates the input token sequences $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ into a concatenated input tensor. The layer normalization block 210 is coupled to receive the concatenated input tensor and normalize the elements of each request to generate a

normalized tensor as the output tensor. The QKV operation block 215 is coupled to receive the normalized tensor as an input tensor and generate an output tensor including queries, keys, values for the requests. Specifically, the QKV operation block 215 generates the queries, keys, and values by applying a QKV weight tensor that is a trained set of parameters of the transformer model 200 to the normalized tensor that includes normalized data for each request to generate an output tensor that includes the queries, keys, and values for each request in the batch. For example, as shown in FIG. 2A, the first row of the output tensor of the QKV operation 215 includes the query, key, value (in a concatenated form) for the first request, the second row includes the query, key, value for the second request, and so on. In one embodiment, instead of applying a single QKV weight tensor, the QKV operation block 215 may be associated with separate weight tensors that are each applied to the input tensor to separately generate the queries, keys, and values. For example, the QKV operation block 215 may apply a query weight tensor to the input tensor to generate the queries for the requests, a separate key weight tensor to the input tensor to generate the keys for the requests, and a separate value weight tensor to the input tensor to generate the values for the requests.

[0041] The split block 220 is coupled to receive the output tensor from the QKV operation block 215 and split the output tensor into a query tensor, a key tensor, a value tensor for the current iteration, in the case that these tensors are not generated separately. The query tensor includes the queries for the batch of requests, the key tensor includes the keys for the batch of requests, and the value tensor includes the values for the batch of requests for the current iteration. In one instance, the execution engine maintains an internal state for the batch of requests as a key cache tensor for caching the keys generated at previous iterations and the current iteration, and a value cache tensor for caching the values that were generated at previous

iterations and the current iteration. The inference system 130 adds the key tensor for the current iteration to the key cache tensor and the value tensor for the current iteration to the value cache tensor.

[0042] The self-attention block 225 is coupled to receive the query tensor, the key cache tensor, and the value cache tensor as the input tensors and generate an output tensor including attention outputs for requests in the batch. In one instance, the attention output tensor is generated by multiplying the query tensor with the key cache tensor to generate a multiplied tensor and multiplying the multiplied tensor with the value cache tensor. When the self-attention block 225 is executed as a batch operation, the query tensor and the key cache tensor may be multiplied on the execution engine to generate the multiplied tensor. The value cache tensor and the multiplied tensor may be multiplied to generate the attention output tensor. The attention output tensor includes attention outputs for the requests in the batch. For example, in FIG. 2A, the first row of the attention output tensor $\mathbf{Z}'$ is the attention output for the first request, the second row of the attention output tensor $\mathbf{Z}'$ is the attention output for the second request, and so on.

[0043] The attention linear block 230 is coupled to receive the attention output tensor as the input tensor and generates an output tensor by applying an attention weight tensor that is a trained set of parameters of the transformer model 200 to the attention output tensor. The attention linear block 230 is configured as a batch operation. For example, in FIG. 2A, the first row of the output tensor $\mathbf{Z}$ may be the output data for the first request, and the second row of the output tensor $\mathbf{Z}$ may be the output data for the second request, and so on. The first add block 235 combines the concatenated input tensor with the input token sequences and the output tensor from the attention linear block 230 to generate a combined tensor. The second layer

normalization block 240 is coupled to receive the combined tensor from the add block 235 and normalize elements of each request to generate a normalized tensor as the output tensor.

[0044] The decoder further includes a multi-layer perceptron (MLP) block 242 that includes one or more blocks including additional neural network layers. In one embodiment, the MLP block 242 includes a first MLP block 245 coupled to receive the normalized tensor from the second layer normalization block 240 as the input tensor and generate an output tensor by applying a MLP weight tensor that is a trained set of parameters of the transformer model 200 to the normalized tensor. The MLP block 242 includes a GeLU block 250 coupled to receive the output tensor from the first MLP block 245 as the input tensor and generate an output tensor by applying a GeLU function to the input tensor. The MLP block 242 includes a second MLP block 255 coupled to receive the output tensor from the GeLU block 250 as the input tensor and generate an output tensor by applying a second MLP weight tensor that is a trained set of parameters of the transformer model 200.

[0045] The combined tensor from the second add block 260 may be obtained as the set of output representations generated for the first decoder D1. Subsequently, the second decoder D2 placed after the first decoder D1 is coupled to receive the set of output representations from the first decoder D1 and repeats similar operations to generate another set of output representations. This process is repeated until the set of output representations from the final decoder DN are generated. Specifically, while each decoder may involve similar operations as the first decoder D1, the trained set of parameters that are associated with the operations may be different from decoder to decoder. The LM head block 270 is coupled to receive the set of output representations from the final decoder DN as an input tensor and generates an output tensor $\hat{\mathbf{y}}_1$ including the one or more output tokens.

[0046] As described in conjunction with FIG. 2B, the next iteration is a decoding phase in which the output tokens generated at the previous iteration are inputs to the transformer model 200. The transformer model 200 generates another set of output tokens. The layer normalization block 210 is applied to the output tensor $\hat{y}_1$ to generate a normalized tensor. The QKV operation block 215 is applied to the normalized tensor to generate an output tensor including the queries, keys, values for the current iteration. The split block 220 is applied to split the output tensor into a query tensor, a key tensor, a value tensor. The inference system 130 adds the key tensor for the current iteration to the key cache tensor and adds the value tensor for the current iteration to the value cache tensor. Thus, during the iteration shown in FIG. 2B, the key cache tensor **Kcache** includes the keys of the previous iterations and the current key tensor **K**, and the value cache tensor **Vcache** includes the values of the previous iterations and the current value tensor **V**.

[0047] The self-attention block 225 is applied to the query tensor, the key cache tensor, and the value cache tensor for the current iteration to generate the attention output tensor. The attention linear block 230 is applied to the attention output tensor to generate an output tensor. The add block 235 combines the inputs and the output tensor from the attention linear block 230 to generate a combined tensor. The MLP block 242 is applied to the combined tensor to generate the set of output representations for the decoder D1. This process is repeated for the remaining decoders and the LM head 270 to generate an output tensor $\hat{y}_2$ including another set of output tokens, and the transformer model 200 is iteratively applied until the termination criteria is satisfied for each request in the batch.

Transformer Model with Selective Batching

[0048] In many instances, the inputs for requests or the internal states for the requests are

variable in length and may be difficult to batch using, for example, the batching method in FIGs. 2A-2B. For example, the number of input tokens between different NLP requests are highly likely to vary in length depending on the input text and the type of request. For example, the length of a first query request may be different from the length of a second query request since different users synthesize language differently. However, methods of batching transformer models, such as the method described in conjunction with FIGs. 2A-2B, require that the lengths of input tensors for the batch be the same since the input tensors have to be coalesced into a single concatenated tensor for an operation. Thus, if the lengths of input tensors are different for multiple requests, it may not be feasible to process the requests as a batch using the batching method of FIGs. 2A-2B.

[0049] In addition, a workaround addressing this problem is to treat the input token sequences during the encoding phase as if the requests had the same lengths as the input token sequence with the shortest length (i.e., least number of tokens). However, this requires the inference system 130 to discard output tokens generated for requests with longer input lengths at subsequent iterations since the original input tokens received for the request have to be substituted in the respective elements. This results in inefficiency and waste of unnecessary resources because an inference system processes the input token sequences for longer requests separately in multiple iterations instead of a single iteration. Another workaround may be to treat the input token sequences as if the requests had the same lengths as the input token sequence with the longest length (i.e., largest number of tokens). However, this requires the inference system to increase the size of the input tensors to match the request with the longest token sequence (and pad the remaining elements). This results in redundant computation since the inference system processes a larger input tensor throughout the process of executing the

transformer model. For example, in some cases, the number of iterations required to process a batch of requests using the method of FIGs. 2A-2B may be significantly larger than the number of iterations to process each request individually.

[0050] Thus, in one embodiment, the inference system 130 trains and applies a machine-learning transformer model to a batch of requests with variable input length or target length or internal state length using a selective batching method. In the selective batching method, the inference system 130 selectively batches a subset of operations in the transformer model but separately processes requests individually for a subset of operations in the transformer model that are not compatible with batching. In one embodiment, the operation to be processed separately is the attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system 130 can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing problems that occur in workarounds.

[0051] FIGs. 3A-3B illustrate a method of selective batching using a machine-learning transformer model 300, in accordance with an embodiment. In the example shown in FIG. 3A, the inference system 130 receives a batch of requests with input token sequences having different lengths, different from FIG. 2A. In FIG. 3A, the inference system 130 receives a first request as the question “what’s up?,” a second request as the question “hello?,” and a third request as the question “how are you?” each having different lengths of words. The inference system 130 encodes each of the requests as a respective set of input token sequences. The first request is encoded to an input token sequence X_1 , the second request is encoded to an input token sequence X_2 , and the third request is encoded to an input token sequence X_3 .

[0052] The transformer model 300 shown in FIGs. 3A-3B also includes a set of N decoders D1, D2, . . . , DN and a LM head block 370. However, different from the transformer model 200 of FIGs. 2A-2B, a subset of operations in the transformer model 300, specifically operations in the self-attention operation block 325, are executed separately instead of being processed as a batch operation. As shown in FIG. 3A, the transformer model 300 includes a first layer normalization block 310, a QKV operation block 315, a split block 320, a self-attention block 325, an attention linear block 330, a first add block 335, a second layer normalization block 340, a first MLP block 345, a GeLU block 350, a second MLP block 355, and a second add block 360.

[0053] Specifically, in the selective batching method, the inference system 130 may concatenate the input token sequences $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ into a concatenated input tensor. Different from the batching method in FIGs. 2A-2B, the inference system 130 concatenates the input token sequences such that individual input tokens for the batch of requests are concatenated across one dimension, for example, the vertical dimension in FIGs. 3A-3B. When input token sequences with different lengths are input to the transformer model 300, one way to concatenate the sequences into a concatenated tensor with the same dimensionality for the requests in the batch is to concatenate the input tokens for the requests across a single dimension. Thus, the concatenated tensor in FIG. 3A is illustrated as a $\Sigma L_b \times 1$ tensor in FIG. 3A, where ΣL_b denotes the total sum of the token lengths over each request b in the batch, in which the length of input tokens for request b is given by L_b . In practice, the concatenated input tensor may be a $\Sigma L_b \times 1 \times H$ tensor where H is the dimension of an embedding.

[0054] The first layer normalization block 310 is applied to the concatenated input tensor to generate a normalized tensor that normalizes elements of each request. In one instance, the

operation of the first layer normalization block 310 is given by:

$$Y_{i,j} = \frac{X_{i,j} - E[\mathbf{X}_{i,\cdot}]}{\sqrt{\text{Var}[\mathbf{X}_{i,\cdot}] + \epsilon}} \cdot \gamma_j + \beta_j$$

where $\mathbf{X} \in \mathbb{R}^{\Sigma L_b \times H}$ represents the concatenated input tensor and $\mathbf{Y} \in \mathbb{R}^{\Sigma L_b \times H}$ represents the normalized tensors, ϵ is a constant, and $\gamma, \beta \in \mathbb{R}^H$.

[0055] The QKV operation block 315 is applied to the normalized tensor as an input tensor to generate an output tensor including the queries, keys, values for the requests. The QKV operation block can be configured as one or more neural network layers. Specifically, the QKV operation block 315 generates the queries, keys, and values by applying a QKV weight tensor that is a trained set of parameters of the transformer model 300 to the normalized tensor. In one instance, the operation of the QKV operation block 315 is given by:

$$Y_{i,j} = \left(\sum_{k=1}^H X_{i,k} \cdot W_{k,j}^{\text{QKV}} \right) + \text{bias}_j$$

where $\mathbf{X} \in \mathbb{R}^{\Sigma L_b \times H}$ represents the normalized tensor from the first layer normalization block 310, $\mathbf{Y} \in \mathbb{R}^{\Sigma L_b \times 3H}$ represents the output tensor containing the queries, keys, and values, $\mathbf{W}^{\text{QKV}} \in \mathbb{R}^{H \times 3H}$ is the QKV weight tensor, and $\text{bias} \in \mathbb{R}^{3H}$.

[0056] In particular, the QKV operation block 315 in the selective batching method may be configured as a batch operation and the QKV weight tensor is multiplied with the normalized tensor from the layer normalization block 310 to generate the output tensor that includes the queries, keys, and values for the batch of requests. Because of the shape of the concatenated input tensor, the output tensor of the QKV operation 315 may also be arranged such that each query element for the requests are concatenated along one dimension, for example, the vertical dimension, each key element for the requests are concatenated along the one dimension, and

each value element for the requests are concatenated along the one dimension. For example, as shown in FIG. 3A, the first and second rows of the output tensor of the QKV operation 315 corresponds to the query, key, value for the first request, the third row corresponds to the query, key, value for the second request, and the fourth through sixth rows correspond to the query, key, value for the third request. Similarly, the QKV operation block 315 may generate the queries, keys, and values separately by applying a query tensor, a separate key tensor, and a separate value tensor to the input tensor. In one embodiment, the QKV operation block 315 includes multiple attention heads, and the queries, keys, and values are generated for each attention head. In such an instance, when the QKV operation block 315 includes n attention heads, H may be equal to $n \times h$, where h is the dimensionality of the latent space per attention head.

[0057] The split block 320 is coupled to receive the output tensor from the QKV operation block 315 and split the output tensor into a query tensor, a key tensor, a value tensor for the current iteration. In one instance, an operation of the split block 320 is given by:

$$Q_{i,j} = X_{i,j}, \quad K_{i,j} = X_{i,H+j}, \quad V_{i,j} = X_{i,2H+j}$$

where $\mathbf{X} \in \mathbb{R}^{\Sigma_L \times 3H}$ represents the output tensor from the QKV operation block 315, $\mathbf{Q} \in \mathbb{R}^{\Sigma_L \times H}$ represents the query tensor, $\mathbf{K} \in \mathbb{R}^{\Sigma_L \times H}$ represents the key tensor, $\mathbf{V} \in \mathbb{R}^{\Sigma_L \times H}$ represents the value tensor. When the QKV operation block 315 includes multiple attention heads, the query tensor may include n queries for each request, the key tensor may include n keys for each request, and the value tensor may include n values for each request.

[0058] Different from the batching method of FIG. 2A, the split block 320 further splits the query tensor $\mathbf{Q}$ to queries for each request, the key tensor $\mathbf{K}$ to keys for each request, and the value tensor $\mathbf{V}$ to values for each request. In one instance the per-request split operation of the split block 320 is given by:

$$Q_{1\ i,j} = Q_{i,j}, \quad Q_{2\ i,j} = Q_{L_1+i,j}, \quad \dots \quad Q_{B\ i,j} = Q_{L_1+\dots+L_{B-1}+i,j}$$

where $\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_B$ are queries for individual requests, and $\mathbf{Q}_1 \in \mathbb{R}^{L_1 \times H}$, $\mathbf{Q}_2 \in \mathbb{R}^{L_2 \times H}, \dots, \mathbf{Q}_B \in \mathbb{R}^{L_B \times H}$,

$$K_{1\ i,j} = K_{i,j}, \quad K_{2\ i,j} = K_{L_1+i,j}, \quad \dots \quad K_{B\ i,j} = K_{L_1+\dots+L_{B-1}+i,j}$$

where $\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_B$ are keys for individual requests, and $\mathbf{K}_1 \in \mathbb{R}^{L_1 \times H}$, $\mathbf{K}_2 \in \mathbb{R}^{L_2 \times H}, \dots, \mathbf{K}_B \in \mathbb{R}^{L_B \times H}$,

$$V_{1\ i,j} = V_{i,j}, \quad V_{2\ i,j} = V_{L_1+i,j}, \quad \dots \quad V_{B\ i,j} = V_{L_1+\dots+L_{B-1}+i,j}$$

where $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_B$ are queries for individual requests, and $\mathbf{V}_1 \in \mathbb{R}^{L_1 \times H}$, $\mathbf{V}_2 \in \mathbb{R}^{L_2 \times H}, \dots, \mathbf{V}_B \in \mathbb{R}^{L_B \times H}$.

[0059] Moreover, for each request $b = 1, 2, \dots, B$, the inference system 130 maintains an internal state for the request in the form of a key cache tensor $\mathbf{Kcache}_b$ and a value cache tensor $\mathbf{Vcache}_b$, where $\mathbf{Kcache}_b \in \mathbb{R}^{L_{b,curr} \times H}$ and $\mathbf{Vcache}_b \in \mathbb{R}^{L_{b,curr} \times H}$ and $L_{b,curr}$ is the current length of the token sequence (including both the input tokens and output tokens) of the request b . In other embodiments, the internal state for a request may contain different types of information from the keys and values for the request and embodiments are not limited hereto. Thus, $L_{b,curr}$ is equal to L_b for the encoding phase. The inference system 130 adds the key tensor $\mathbf{K}_b$ to the respective key cache tensor and the value tensor $\mathbf{V}_b$ to the respective value cache tensor for the request. Because the set of inputs (i.e., input token sequences for $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$) for the batch of requests have different lengths, the length of the internal state for the requests in the batch are different in FIG. 3A. In the batching process of FIGs. 2A-2B, the key cache tensor and the value cache tensor are also required to have the same lengths across requests in the batch since the self-attention operation 225 involves multiplying the query tensor, the key cache tensor, and the value cache tensor together. However, since selective batching is performed in FIG. 3A,

requests that have different lengths for its internal state can also be efficiently processed in batches for select operations while being processed separately for the self-attention operation 325.

[0060] The self-attention block 325 is coupled to receive the query tensor, the key cache tensor, and the value cache tensor for each request as the one or more input tensors and generates one or more output tensors that are attention outputs for each request in the batch. In one embodiment, the self-attention block 325 is not batched, and at least a subset of the requests in the batch are processed separately from one another. In particular, each request in the batch may be processed individually for the self-attention block 325.

[0061] For a request in the batch, the self-attention block 325 is coupled to receive the query for the request and the key cache tensor for the request to generate a multiplied tensor for the request. In one instance, an operation of the self-attention block 325 for a request is given by:

$$Y_{b\ i,j,k} = \sum_{m=1}^h Q'_{b\ i,j,m} \cdot Kcache'_{b\ i,m,k}$$

where $Y_b \in \mathbb{R}^{n \times L_{b,curr} \times L_{b,curr}}$ represents the multiplied tensor for request b , $Q'_b \in \mathbb{R}^{n \times L_{b,curr} \times h}$ is a reshaped query tensor Q_b for request b that concatenates the query for each attention head together for a request b , and $Kcache'_b \in \mathbb{R}^{n \times h \times L_{b,curr}}$ is a reshaped key cache tensor $Kcache_b$ for request b that concatenates the key cache tensor for each attention head together for request b .

[0062] The self-attention block 325 is then coupled to receive the multiplied tensor and the value cache tensor for the request and generate the attention output tensor for the request. In one instance, an operation of the self-attention block 325 for the request is given by:

$$P_{b\ i,j,k} = \frac{\exp\left(\frac{1}{\sqrt{h}} \cdot X_{b\ i,j,k}\right)}{\sum_{m=1}^j \exp\left(\frac{1}{\sqrt{h}} \cdot X_{b\ i,j,m}\right)} \quad \text{if } k \leq j$$

$$P_{b\ i,j,k} = 0 \quad \text{if } k > j \quad ,$$

Where $\mathbf{X}_b \in \mathbb{R}^{n \times L_{b,curr} \times L_{b,curr}}$ represents the received multiplied tensor for request b , $\mathbf{P}_b \in \mathbb{R}^{n \times L_{b,curr} \times L_{b,curr}}$ and

$$Y_{b\ i,j,k} = \sum_{m=1}^{L_{b,curr}} P_{b\ i,j,m} \cdot V_{cache'}'_{b\ i,m,k}$$

where $\mathbf{Y}_b \in \mathbb{R}^{n \times L_{b,curr} \times h}$ represents the multiplied tensor for request b , $\mathbf{V}_{cache'}'_b \in \mathbb{R}^{n \times L_{b,curr} \times h}$ is the reshaped value tensor $\mathbf{V}_{cache}_b$ for request b that concatenates the value cache tensor for each attention head together for request b . Subsequently the output $\mathbf{Y}_b$ is reshaped to generate the attention output $\mathbf{Z}'_b$ by:

$$Z'_{b\ j,hi+k} = Y_{b\ i,j,k}$$

where $\mathbf{Y}_b \in \mathbb{R}^{n \times L_{b,curr} \times h}$ and $\mathbf{Z}'_b \in \mathbb{R}^{L_{b,curr} \times H}$.

[0063] In particular, the operations of the self-attention block 325 may not be compatible as a batch operation for requests that have different lengths because the operations of the self-attention block 325 are multiplications between the queries, key caches, and value caches for the requests rather than tensor multiplication between a fixed weight tensor and one or more input tensors. Thus, the attention outputs for requests in the batch are generated separately. For example, an execution engine may generate a first attention output for a first request using input tensors (e.g., query tensor $\mathbf{Q}_1$, key cache tensor $\mathbf{K}_{cache}_1$, value cache tensor $\mathbf{V}_{cache}_1$) for the first request, and separately, the execution engine at a separate step may generate a second attention output for a second request using input tensors for the second request. In particular, the second attention output may be generated at a different execution engine from the execution engine that processed the first request, generated at a different hardware accelerator from the hardware accelerator that processed the first request in the same execution engine, generated at a

different GPU kernel from the GPU kernel that processed the first request in the same hardware accelerator, or generated at the same GPU kernel that processed the first request in the same hardware accelerator, but embodiments are not limited hereto.

[0064] The inference system 130 concatenates the attention outputs of the requests in the batch together into a single attention output tensor, since the subsequent operation in the attention linear block 330 is configured as a batch operation. In one instance, the concatenation operation is given by:

$$\mathbf{Z}'_{i,j} = \mathbf{Z}'_{1\ i,j}, \quad \mathbf{Z}'_{L_1+i,j} = \mathbf{Z}'_{2\ i,j}, \quad \dots \quad \mathbf{Z}'_{L_1+\dots+L_{B-1}+i,j} = \mathbf{Z}'_{B\ i,j}$$

where $\mathbf{Z}' \in \mathbb{R}^{\Sigma L_b \times H}$ represents the single attention output tensor. As shown in FIG. 3A, the first and second rows of the concatenated attention output tensor $\mathbf{Z}'$ correspond to attention outputs of the first request, the third row corresponds to the attention outputs of the second request, and the fourth through sixth rows correspond to the attention outputs of the third request.

[0065] The attention linear block 330 is coupled to receive the attention output tensor as the input tensor and generates an output tensor by applying an attention weight tensor that is a trained set of parameters of the transformer model 300 to the attention output tensor. The attention linear block 330 may be configured as one or more neural network layers. The attention linear block 330 may be executed as a batch operation. In one instance, the operation of the attention linear block 330 is given by:

$$Z_{i,j} = \left(\sum_{k=1}^H \mathbf{Z}'_{i,k} \cdot \mathbf{W}_{k,j}^{\text{attn}} \right) + \text{bias}_j$$

where $\mathbf{Z}' \in \mathbb{R}^{\Sigma L_b \times H}$ represents the attention output tensor from the self-attention block 325, $\mathbf{Z} \in \mathbb{R}^{\Sigma L_b \times H}$ represents the output tensor, $\mathbf{W}^{\text{attn}} \in \mathbb{R}^{H \times H}$ is the attention weight tensor, and $\text{bias} \in \mathbb{R}^H$.

As an example, in FIG. 3A, the first to second rows of the output tensor $\mathbf{Z}$ include the output data

for the first request, the third row includes the output data for the second request, and the fourth to sixth rows includes the output data for the third request.

[0066] The first add block 335 combines the concatenated input tensor including the input token sequences and the output tensor from the attention linear block 330 to generate a combined tensor. The second layer normalization block 340 is coupled to receive the combined tensor from the add block 335 and normalize elements of each request to generate a normalized tensor as the output tensor.

[0067] Similar to the transformer model 200 of FIGs. 2A-2B, the decoder may further include a MLP block 342 that includes one or more blocks including additional neural network layers. The MLP block 342 is executed as one or more batch operations. The MLP block 342 further includes a first MLP block 345 configured as one or more neural network layers. The first MLP block 345 is coupled to receive the normalized tensor from the second layer normalization block 340 and generate an output tensor by applying a MLP weight tensor that is a trained set of parameters of the transformer model 200 to the normalized tensor.

[0068] The MLP block 342 further includes a GeLU block 350 coupled to receive the output tensor from the first MLP block 345 as the input tensor and generate an output tensor by applying a GeLU function to the input tensor. The MLP block 342 also includes a second MLP block 355 configured as one or more neural network layers. The second MLP block 355 is coupled to receive the output tensor from the GeLU block 350 as the input tensor and generate an output tensor by applying a second MLP weight tensor that is a trained set of parameters of the transformer model 300. The second add block 360 combines the output tensor from the second MLP block 355 and the output tensor from the first add block 335 to form the combined tensor.

[0069] The combined tensor includes the set of output representations for the first decoder D1. The set of output representations are propagated to subsequent decoders and a similar process is repeated to generate a set of output representations for the final decoder DN. Specifically, the inference system 130 may also deploy a selective batching method for the remaining decoders in which the self-attention operation is executed without batching but the remaining operations are batched. The LM head block 370 is coupled to receive the set of output representations from the final decoder DN and generate an output tensor $\hat{\mathbf{y}}_1$ including a set of output tokens.

[0070] As shown in FIG. 3B, the next iteration is a decoding phase where the previous output tokens generated at the previous iteration are inputs to the transformer model 300 to generate another set of output tokens. Specifically, the layer normalization block 310 is applied to the output tensor $\hat{\mathbf{y}}_1$ to generate a normalized tensor. The QKV operation block 315 is applied to the normalized tensor to generate the output tensor including the queries, keys, values for the current iteration. The split block 320 is applied to split the output tensor into one or more query tensors for each request, one or more key tensors for each request, and one or more value tensors for each request in the batch. The inference system 130 adds each key tensor to the key cache tensor for the respective request and adds each value tensor to the value cache for the respective request. Since FIG. 3B depicts an iteration of a decoding phase for all requests in the batch, the requests have the same lengths for the set of inputs (i.e., single output token for each request generated from the previous iteration) that are fed to the transformer model 300. However, the internal state including the key cache tensor and value cache tensor may still have different lengths for requests in the batch, since the input token lengths for the requests were different in the encoding phase illustrated in FIG. 3A. However, similar to FIG. 3A, the

inference system 130 can still process a batch of requests that have different internal state lengths by selectively batching operations in the transformer model 300.

[0071] The self-attention block 325 is separately applied to the query tensor, the key cache tensor, and the value cache tensor for each respective request to generate the attention output tensor for the request. The attention outputs for the requests are concatenated into the attention output tensor. The attention linear block 330 is applied to the attention output tensor to generate an output tensor. The add block 335 combines the concatenated input tensor and the output tensor from the attention linear block 330 to generate a combined tensor. The MLP block 342 is applied to the combined tensor to generate the set of output representations for the decoder D1. This process is repeated for the remaining decoders and the LM head block 370 to generate an output tensor $\hat{y}_2$ including another set of output tokens, and the transformer model 300 is iteratively applied until the termination criteria is satisfied for each request in the batch.

[0072] By performing selective batching on select operations of the transformer model 300, the inference system 130 can efficiently process a batch of requests with variable input length. The inference system 130 executes requests by separately processing operations, such as the self-attention operation, on a per-request basis while remaining operations of the transformer model 300 are batched. Specifically, operations that involve tensor multiplication between one or more input tensors and a trained weight tensor, such as the QKV linear operation, the attention linear operation or the MLP operation, may require high computational power during the inference process since the dimensionality of the weight tensors are often large. Moreover, when the requests are individually processed for these operations, the weight tensors may be read from the local memory of an execution engine multiple times. However, by configuring these subsets of operations as batch operations but separately processing operations, such as the self-

attention operation, that do not involve operations on weight tensors, the inference system 130 can utilize the parallel computation capabilities of hardware accelerators while allowing high flexibility in processing variable length requests.

[0073] In addition, FIG. 3A illustrates an iteration of the encoding phase for a set of inputs that are input token sequences and FIG. 3B illustrates an iteration of the decoding phase for a set of inputs that are output tokens for the batch of requests that were generated in a previous iteration (i.e., encoding phase). However, different from other methods of batching transformer models, the selective batching method described herein flexibly allows the transformer model 300 to process a set of inputs for a batch of requests having variable input lengths, target lengths, or internal state lengths, regardless of whether a request in the batch is processed for an encoding phase or a decoding phase or how many iterations have been processed for the request.

[0074] Specifically, a batch of requests may not be batched using the method of FIGs. 2A-2B when (i) the requests are in the encoding phase and have input token sequences with different lengths, when (ii) the requests are in the decoding phase and each request is processing a token at a different index (i.e., input token length plus the index of decoding iterations) from each other resulting in internal states with different lengths, and when (iii) each request in the batch is in a different phase (encoding or decoding) resulting in inputs with different lengths (i.e., input token sequence for encoding phase and single output token for decoding phase). While the method of batching in FIGs. 2A-2B restrict the requests in a batch to the same phase with the same number of input tokens for the encoding phase and the same token index for the decoding phase, this restriction significantly reduces the likelihood of batching in real-world workloads.

[0075] In contrast, the selective batching method described herein allows the transformer model 300 to process requests as a batch, even if they are at different phases or different indices

for processing. Thus, for example, instead of receiving a set of input token sequences X_1 , X_2 , X_3 , in FIG. 3A, the transformer model 300 may perform an iteration using a batch of an input token sequence for a first request (encoding phase), a first output token for a second request (first iteration of decoding phase), and a third output token for a third request (third iteration of decoding phase). As another example, the transformer model 300 may perform an iteration using a batch of a second output token for a first request and a fourth output token for a second request, each having different lengths for the key cache tensor and the value cache tensor. As described in conjunction with FIGs. 5A-5B, this allows the inference system 130 to flexibly modify and update batches at each iteration so that the computing capabilities of the one or more execution engines are fully utilized.

[0076] Moreover, while FIGs. 2A-2B and 3A-3B illustrate a GPT (generative pre-training) type transformer model that includes a set of decoders, the inference system 130 is also capable of performing selective batching on transformer models with other types of architectures, as long as the transformer model includes a attention mechanism (e.g., self-attention or encoder-decoder attention) that generates attention outputs using queries, keys, and values generated for the request, and generates output tokens in an auto-regressive manner by using the output tokens generated at previous iterations as inputs for the next iteration.

[0077] In another embodiment, a transformer model may have an encoder-decoder architecture and includes a set of encoders coupled to a set of decoders. In such an architecture, each encoder includes at least a self-attention block coupled to a MLP block. The self-attention block of an encoder is coupled to receive a query tensor, a key tensor, and a value tensor obtained from processing a set of input tokens and generate an attention output. Each decoder includes at least a self-attention block coupled to an encoder-decoder attention block coupled to a

MLP block. The encoder-decoder attention block of a decoder is coupled to receive a query tensor obtained from processing the output of the self-attention block and a key tensor and a value tensor obtained from processing the output of the final encoder to generate another attention output.

[0078] The inference system 130 may apply the encoder-decoder transformer model to a batch of requests. Specifically, for an encoder, the requests may be separately processed for the self-attention block while they are batched for the remaining operations. For the decoder, the requests may be separately processed for the self-attention block and the encoder-decoder attention block while they are batched for the remaining operations.

[0079] Returning to FIG. 1, the client devices 110A, 110B is a computing device such as a smartphone with an operating system such as ANDROID® or APPLE® IOS®, a tablet computer, a laptop computer, a desktop computer, or any other type of network-enabled device. In the system environment 100 shown in FIG. 1, the client device 110 allows a user to submit requests to the inference system 130 to perform tasks that can be processed by the machine-learning transformer model. For example, a user of the client device 110A may be affiliated with an entity that deploys one or more applications that require NLP or other types of tasks that can be processed by the transformer architecture. A request submitted by the client device 110 may be, for example, a request to generate predictions for chatbot applications, a request to translate a sequence of words from one language to another language, a request to synthesize a story or narrative given a sequence of starting words, and the like.

[0080] A typical client device 110 includes hardware and software needed to connect to the network 122 (e.g., via WiFi and/or 4G, 5G or other wireless telecommunication standards). Specifically, the client device 110 may include an operating system and various applications that

run on the operating system that enable the users to submit the requests. For example, the client device 110 may include browser applications or standalone applications deployed by the inference system 130 that allow users of an organization to interact with the inference system 130 to submit the requests.

[0081] The network 122 provides a communication infrastructure between the client devices 110 and the online system 130. The network 122 is typically the Internet, but may be any network, including but not limited to a Local Area Network (LAN), a Metropolitan Area Network (MAN), a Wide Area Network (WAN), a mobile wired or wireless network, a private network, or a virtual private network.

BLOCK DIAGRAM OF INFERENCE SYSTEM

[0082] FIG. 4 is a block diagram of an architecture of the inference system 130, in accordance with an embodiment. The inference system 130 shown by FIG. 4 includes a data management module 420, a training module 430, and a serving system 435. The inference system 130 also includes a training corpus 460. In alternative configurations, different and/or additional components may be included in the inference system 130.

[0083] The data management module 420 manages the training corpus 460 of training data that is used to train parameters of the transformer model. In one embodiment, the training corpus 460 includes multiple instances of data that each include a set of training input data and a set of training output data that correspond to known predictions for the set of training input data that address the task the transformer model should be trained for. In one embodiment, the training input data and the training output data is text data, but embodiments are not limited hereto, and the training data may include any type of data, such as audio data or image data, that the transformer model is trained to process in an auto-regressive manner. The training module

430 encodes the set of training input data to a set of training input tokens and the set of training output data to a set of training output tokens. Each token may represent a respective word in the latent space.

[0084] For example, when the transformer model is used for automatic translation from English to French, the set of training input tokens may correspond to a first sentence of words in English and the set of training output tokens may correspond to a second sentence of words in French that is a translation of the first sentence. As yet another example, when the transformer model is used to convert an image to text, the set of training input tokens may correspond to pieces of image data for the training image and the set of training output tokens may correspond to a converted sentence or paragraph describing the image.

[0085] The execution engine module 425 includes or more execution engines that are built on specialized hardware accelerators such as graphics processing units (GPU's) or tensor processing units (TPU's). An execution engine receives requests to execute one or more requests on the specialized hardware. In one instance, the execution engine receives a batch of requests and executes one or more iterations of the transformer model via selective batching using the inputs for each request. As described above, an execution engine may include a set of cores (e.g., GPU cores) coupled to local memory (e.g., GPU memory). Specifically, for each operation of an iteration, the execution engine is configured to read data required for the operation from the local memory of the execution engine.

[0086] The execution engines of the execution engine module 425 may be configured as GPU's or TPU's or any other specialized hardware that is capable of processing neural network operations, including tensor multiplication and floating-point operations in an efficient manner. For example, the execution engine is not limited to Google TPU, Graphcore IPU, Cerebras WSE,

Nvidia GPU, intel Nervana, Qualcomm Cloud AI, Samsung Exynos, AMD Radeon, Xilinx AI Accelerator cards, IBM TrueNorth, AWS Trainium, and the like. As described above, an execution engine may be composed of one or more hardware accelerators. In some instances, the execution engine may process workload that requires processing capabilities of more than one hardware accelerator. Thus, the execution engine is able to distribute the workload across multiple hardware accelerators if necessary.

[0087] In one embodiment, an execution engine is also configured to manage one or more caches on the local memory necessary for executing one or more iterations of the transformer model. For example, when executing the transformer model 300 of FIGs. 3A-3B, the execution engine maintains a key cache tensor and a value cache tensor for a request until the request has been completed. Thus, responsive to receiving one or more new requests (requests for which the encoding phase has not been processed), an execution engine allocates memory to each request for maintaining the key cache tensor and the value cache tensor for the request. For each iteration of the decoding phase, the execution engine may add the key tensor and the value tensor for the iteration to the respective key cache and value cache allocated for the request and retrieve the key cache and value cache for the request to execute an operation. Responsive to completing one or more requests, an execution engine may provide the outputs for the requests to an appropriate module of the inference system 130, and free the allocated cache memory for the completed requests, such that the freed memory can be used for other requests.

[0088] The training module 430 trains parameters of a transformer model by performing a training process. First, the training module 430 may generate an appropriate architecture for the transformer model that can address the task requests received by the client devices 110. For example, the training module 430 may train the transformer model 300 illustrated in FIGs. 3A-

3B . The training module 430 may determine the number of decoders in the set of decoders to adjust the depth of the transformer model and determine, for example, how many layers are included in each neural network layer of the transformer model. When the transformer model to be trained is an encoder-decoder architecture, the training module 430 may determine the number of encoders as well as the number of decoders.

[0089] The training module 430 may also initialize one or more weight tensors associated with the operations of the transformer model. The elements of the weight tensors correspond to parameters of the transformer model that will be learned during the training process using the training data of the training corpus 460. The weight tensors may be initialized with dimensionality based on the dimensionality of the input tensors the weight tensors are configured to be operated with. For example, to train the transformer model 300 of FIGs. 3A-3B, the training module 430 may initialize a QKV weight tensor $\mathbf{W}^{\text{QKV}}$ for the QKV operation block 315 (e.g., with dimensionality $H \times 3H$), an attention weight tensor $\mathbf{W}^{\text{attn}}$ for the attention linear block 330 (e.g., dimensionality $H \times H$), and a first MLP weight tensor for the first MLP block 345 and a second MLP weight tensor for the second MLP block 355.

[0090] During the training process, the training module 430 obtains a set of training data and trains parameters of the machine-learning transformer model by repeatedly iterating between a forward pass step and a backpropagation step. During the forward pass step, the training module 430 generates one or more estimated output tokens by applying estimated parameters of the transformer model to the set of training input tokens in the set of training data to generate one or more estimated output tokens. The training module 430 determines a loss function indicating a difference between the one or more estimated output tokens and the set of training output tokens. During the backpropagation step, the training module 430 updates parameters of

the transformer model (i.e., weight tensors) to reduce the loss function. This process is iteratively repeated for the next sets of training data until convergence is reached for parameters of the transformer model or a desired level of model quality is reached.

[0091] In one embodiment, the training module 430 arranges the training data in batches and executes one or more iterations of the training process on one or more execution engines. For example, each batch of training data may include a batch of input token sequences, each from a different training source. For example, a first training input token sequence may be obtained from a first text source and a second training input token sequence may be obtained from a second text source. By executing the training process with batching, the training module 430 can utilize the parallel processing capabilities of the execution engines when training the transformer model, which requires a high degree of computation.

[0092] The serving system 435 receives requests from client devices 110 to perform one or more tasks that can be processed using the trained transformer model. As described with respect to FIGs. 3A-3B, when the tasks are NLP applications, the requests may include a sequence of words (e.g., sentence in English) that the serving system 435 can parse and map into a sequence of input tokens that can be provided to the transformer model. The serving system 435 processes the request to generate one or more output tokens using the transformer model. The serving system 435 can convert the output tokens to output data, for example, a sequence of words (e.g., translated sentence in French), and return the output data as a response to the client device.

[0093] In one embodiment, the serving system 435 receives multiple requests from client devices 110 over time and forms batches of requests. The serving system 435 provides the batches to one or more execution engines. Responsive to execution by the execution engines,

the serving system 435 receives one or more output tokens for the requests and provides the outputs to the client devices 110 as response to the requests. In one embodiment, the lengths of input sequences of the requests in a batch or lengths of the internal states of the requests in a batch formed by the serving system 435 can be different from one another as the execution engines are capable of processing the requests using the selective batching method. In one embodiment, the serving system 435 waits until a request is completed (i.e., the termination criteria is satisfied), and provides the outputs to the client devices 110 as response to the requests. In another embodiment, the serving system 435 provides outputs to the client devices 110 even though a request has not been completed yet. For example, the serving system 435 may provide a number (e.g., 5, 10, 15) of output tokens every time that the number of output tokens is generated for the request. As another example, for one or more time intervals (e.g., 100 ms, 200 ms) the serving system 435 may provide the output tokens generated after each time interval to the client device 110. This allows the user of the client device 110 to receive intermittent outputs even though a request has not been completed yet.

[0094] In one embodiment, the serving system 435 performs iteration-level dynamic batching for the transformer model that allows the serving system 435 to dynamically modify a batch of requests being executed on an execution engine on a per-iteration basis. Specifically, it may be difficult to modify a batch of requests once processing has started on an execution engine for certain batching methods because such methods require the length of inputs or the lengths of internal states to the transformer model be the same across the requests of the batch. That is, unless new incoming requests have the same length of inputs and internal states as the batch of requests being executed, it is difficult for the inference system to modify the batch to, for example, add new requests to the batch.

[0095] By performing selective batching, the serving system 435 can monitor and modify a batch processed on the execution engine per iteration. Specifically, at one or more iterations, the serving system 435 can modify the batch being executed on the execution engine by adding new incoming requests to the batch or removing completed requests from the batch. This is because selective batching allows requests with variable lengths to be processed without restraining the inputs or the internal states to the transformer model to the same lengths. This allows the serving system 435 to provide the response for completed requests earlier to the client device 110 of the request and allows addition of new requests to a batch if the execution engine processing the batch is being under-utilized, even if the lengths for the new requests are different from lengths of the existing requests for the next iteration.

Dynamic Batching for Transformer Model

[0096] FIGs. 5A-5D illustrate a method of dynamic batching for processing requests using a machine-learning transformer model, in accordance with an embodiment. In one embodiment, the serving system 435 includes a request processor 580 and a scheduler 585 each coupled to the one or more execution engines. The request processor 580 receives requests and forwards the requests to the scheduler 585. The request processor 580 maintains a completion queue for storing outputs of completed requests. Specifically, the request processor 580 receives outputs for completed requests from the execution engines and stores the outputs in the completion queue such that they can be provided to the client devices 110 of the requests. The scheduler 585 receives the forwarded requests from the request processor 580 and maintains an incoming request queue for storing new requests to be processed. The scheduler 585 forms a batch of requests and schedules the batch for execution on an execution engine. In one embodiment, the scheduler 585 is configured to monitor which batch of requests were distributed to each

execution engine and how many iterations of the transformer model have been performed for each request and whether the request has been completed. The scheduler 585 is also configured to monitor the available cache memory in each execution engine.

[0097] Specifically, FIG. 5A illustrates a request processor 580 and a scheduler 585 coupled to execution engines 590A and 590B. In FIG. 5A, execution engine 590A is scheduled to execute a single request R1, and execution engine 590B is scheduled to execute a batch of requests R3, R4, R5. Specifically, for the first iteration of FIG. 5A, request R1 in execution engine 590A includes a single input token. On the other hand, request R3 in the execution engine 590B includes a sequence of two input tokens, request R4 includes a sequence of three input tokens, and request R5 includes a sequence of two input tokens.

[0098] The request processor 580 receives a new request R2 and forwards request R2 to the scheduler 585. The scheduler 585 stores request R2 in the incoming request queue. An iteration of the transformer model is executed on the execution engines 590A, 590B. For example, the iteration may be an encoding phase for a certain request in current executing batch. In particular, a key cache tensor and value cache tensor are allocated for each request as the internal state cache, and a part of the cache for each request is used to store the keys and values after the encoding phase. In one embodiment, the allocation of the internal state cache for a request is based on the maximum output token length the transformer model is configured to generate. For example, a transformer model may be configured to generate a maximum number of 1024 output tokens, and the execution engine may allocate an internal state cache configured to store up to 1024 elements for each key cache tensor and value cache tensor. In another embodiment, when a user of the request specifies a maximum length of output tokens that can be returned as the response, and the execution engine may allocate an internal state cache

configured to store up to the maximum length of elements specified in the request for each tensor. In yet another embodiment, the execution engine may determine the initial number of input tokens for the request, and the execution engine may initially allocate an internal state cache that is large enough to store the number of input tokens plus some additional amount of memory. The execution engine may incrementally allocate more memory to the internal state cache as the transformer model is iteratively applied for the request if the initially allocated memory is insufficient. The scheduler 585 monitors the cache memory for execution engines 590A and 590B. Responsive to determining that execution engine 590A has cache memory available for processing request R2, the scheduler 585 updates the batch for execution engine 590A to include the new request R2 and instructs the execution engine 590A to execute the updated batch for the next iteration.

[0099] As shown in FIG. 5B, after executing the first iteration, a first output token is generated for request R1 in the execution engine 590A and requests R3, R4, R5 in the execution engine 590B. Moreover, execution engine 590A is now scheduled to execute an updated batch of requests R1, R2 at the second iteration. The request R2 includes a single input token. In particular, since the first output token has been generated for request R1, the second iteration for request R1 may be a decoding phase, while the second iteration for request R2 may be an encoding phase. Thus, because an execution engine is configured to perform selective batching and is capable of processing requests with different lengths, execution engine 590A can perform both the encoding phase and the decoding phase for the same batch of requests, as described in conjunction with FIGs. 3A-3B. This may not be feasible for other batching methods as there is no guarantee that the length of input tokens for one request will be the same as the length of input tokens and output tokens for an existing request in the batch. As shown in FIG. 5B, an

additional part of the cache for each request is used to store the keys and values after the second iteration.

[00100] As shown in FIG. 5C, after executing the second iteration, a second output token is generated for request R1 in execution engine 590A and requests R3, R4, R5 in the execution engine 590B. A first output token is generated for request R2 in execution engine 590A. Specifically, the first output token generated for request R2 is generated with an end token and the execution engine 590A provides the outputs for request R2 to the completion queue of the request processor 580. The execution engine 590A frees the cache memory allocated to request R2. The second output token generated for request R4 is also generated with an end token and the execution engine 590B provides the outputs for request R4 to the completion queue of the request processor 580. The execution engine 590B frees the cache memory allocated to request R4. The request processor 580 forwards the outputs of the completion queue to the client devices 110 of the completed requests.

[00101] The request processor 580 also receives another new request R7 and forwards the request to the scheduler 585. The scheduler 585 stores the request R7 in the incoming request queue. Responsive to determining that requests R2, R4 are completed and that execution engine 590A has cache memory available for processing request R7, the scheduler 585 updates the batch for execution engine 590A to R1, R7 and updates the batch for execution engine 590B to R3, R5. The scheduler 585 instructs the execution engines 590A, 590B to execute the updated batches. Thus, by dynamically adjusting the batches at an iteration, completed requests can be provided to the client devices 110 of the requests as soon as processing is completed, and the scheduler 585 can schedule new requests within the batch such that the computing capabilities of an execution engine can be utilized when the memory of the completed request is

freed up.

[00102] As shown in FIG. 5D, after executing the third iteration, a third output token is generated for request R1 in the execution engine 590A and requests R3, R5 in the execution engine 590B. Moreover, execution engine 590A is now scheduled to execute an updated batch of requests R1, R7 at the fourth iteration and execution engine 590B is now scheduled to execute an updated batch of requests R3, R5 at the fourth iteration. The request R7 includes a sequence of two input tokens. A similar process to that described in FIGs. 5A-5D may be continuously performed as the serving system 435 receives requests from client devices 110.

[00103] As described above, dynamic batching allows the serving system 435 to dynamically adjust batches that are processed on the execution engines such that the hardware of the execution engine can be fully utilized. This process was not available for certain methods of batching (e.g., method of FIGs. 2A-2B) for transformer models, because internal states, such as the key cache tensor and value cache tensor, maintain variable length per iteration in a transformer model, while other recurrent machine-learning models (e.g., RNN's) make use of state data that have the same lengths across a batch of requests.

FLOWCHART ILLUSTRATING METHOD OF SELECTIVE BATCHING AND DYNAMIC BATCHING

[00104] FIGs. 6A-6B is a flowchart illustrating a method of selective batching using the transformer model, in accordance with an embodiment. The inference system 130 receives 602 a batch of requests including one or more input token sequences. In one instance, a length of a first token sequence for a first request in the batch may be different from a length of a second token sequence for a second request in the batch. The inference system 130 accesses 604 a machine-learning transformer model including at least a set of decoders.

[00105] For one or more iterations, the inference system 130 repeatedly performs the steps of

applying the set of decoders to one or more inputs for the requests. Specifically, for at least one decoder in the set, the inference system 130 generates 606 one or more queries, one or more keys, and one or more values for the requests by applying a QKV weight tensor to one or more input representations. The queries, keys, and values may be generated by a batch operation. The inference system 130 splits 608 a first query for the first request from the one or more queries, a first key for the first request from the one or more keys, and a first value for the first request from the one or more values. The inference system 130 also splits a second query for the second request from the one or more queries, a second key for the second request from the one or more keys, and a second value for the second request from the one or more values.

[00106] The inference system 130 generates 610 a first attention output for the first request by at least combining the first query, the first key, and the first value. The inference system 130 separately generates 612 a second attention output for the second request by combining the second query, the second key, and the second value for the second request. The inference system concatenates 614 at least the first attention output and the second attention output into a concatenated tensor. The inference system 130 generates 616 one or more output representations by applying a weight tensor to the concatenated tensor. The one or more output representations may be generated by a batch operation. The inference system 130 sets 618 the one or more output tokens as the one or more inputs to the set of decoders for a next iteration. The inference system 130 provides 620 output tokens generated for at least one request to a client device 110 as a response to the at least one request.

[00107] FIG. 7 is a flowchart illustrating a method of dynamic batching for processing requests using the transformer model, in accordance with an embodiment. The inference system 130 receives 710, by a serving system, one or more requests for execution. The serving

system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders. The inference system 130 schedules 712, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine. The inference system 130 generates 714, by the execution engine, a first set of output tokens by iteratively applying the transformer model to a first set of inputs for the batch of requests. Applying the transformer model includes applying at least one batch operation to one or more input tensors associated with the batch of requests.

[00108] The inference system 130 receives 716, by a request processor, a new request from a client device. The new request may include a sequence of input tokens. The inference system 130 schedules 718, by the scheduler, a second batch of requests including the one or more requests and the new request for execution on the execution engine responsive to determining that the execution engine has memory available to execute the second batch of requests. The inference system 130 generates 720, by the execution engine, a second set of output tokens by iteratively applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request.

HARDWARE COMPONENTS

[00109] FIG. 8 is a diagram illustrating a computer system 800 upon which embodiments described herein may be implemented within the inference system 130. For example, in the context of FIG. 1, the inference system 130 and its sub-systems may be implemented using a computer system such as described by FIG. 8. The inference system 130 and its sub-systems may also be implemented using a combination of multiple computer systems as described by FIG. 8.

[00110] In one implementation, the computer system 800 includes processing resources 801,

main memory 803, read only memory (ROM) 805, storage device 807, a communication interface 809, and hardware accelerators 810. The computer system 800 includes at least one processor 801 including CPU cores for processing information and a main memory 803, such as a random-access memory (RAM) or other dynamic storage device, for storing information and instructions to be executed by the processor 801. Main memory 803 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 801. The computer system 800 may also include ROM 805 or other static storage device for storing static information and instructions for processor 801. The storage device 807, such as a magnetic disk or optical disk or solid-state memory device, is provided for storing information and instructions. For example, in the context of FIG. 4, the training corpus store 460 of the inference system 130 may be stored in the read only memory (ROM) 805 or the storage device 807. The computer system 800 also includes one or more hardware accelerators 810. As described above, while an execution engine may be configured with the one or more hardware accelerators 810 in a single computer system 800, in another embodiment, an execution engine may be configured across multiple hardware accelerators 810 that are distributed across multiple computer systems 800.

[00111] The communication interface 809 can enable the inference system 130 to communicate with client devices 110 through use of a communication link (wireless or wireline). Using the communication link, the inference system 130 can communicate with the different sub-systems included in a client device 110 to enable receiving and processing of requests. In addition, when an execution engine is configured across multiple hardware accelerators that are distributed across multiple computer systems 800, the communication interface 809 can also enable communication between the multiple computer systems 800, such that an execution

engine can process requests across the multiple hardware accelerators 810.

[00112] The computer system 800 can optionally include a display device 811, such as a cathode ray tube (CRT), an LCD monitor, an LED monitor, a TFT display or a television set, for example, for displaying graphics and information to a user. An input mechanism 813, such as a keyboard that includes alphanumeric keys and other keys, can optionally be coupled to the computer system 800 for communicating information and command selections to processor 801. Other non-limiting, illustrative examples of input mechanisms 813 include a mouse, a trackball, touch-sensitive screen, or cursor direction keys for communicating direction information and command selections to processor 801 and for controlling cursor movement on display device 811.

[00113] Examples described herein are related to the use of the inference system 130 for implementing the techniques described herein. According to one embodiment, those techniques are performed by the inference system 130 in response to processor 801 executing one or more sequences of one or more instructions contained in main memory 803. Such instructions may be read into main memory 803 from another machine-readable medium, such as storage device 807. Execution of the sequences of instructions contained in main memory 803 causes processor 801 to perform the process steps described herein. In alternative implementations, hard-wired circuitry may be used in place of or in combination with software instructions to implement examples described herein. Thus, the examples described are not limited to any specific combination of hardware circuitry and software.

SUMMARY

[00114] The foregoing description of the embodiments of the invention has been presented for the purpose of illustration; it is not intended to be exhaustive or to limit the invention to the

precise forms disclosed. Persons skilled in the relevant art can appreciate that many modifications and variations are possible in light of the above disclosure.

[00115] Some portions of this description describe the embodiments of the invention in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are commonly used by those skilled in the data processing arts to convey the substance of their work effectively to others skilled in the art. These operations, while described functionally, computationally, or logically, are understood to be implemented by computer programs or equivalent electrical circuits, microcode, or the like. Furthermore, it has also proven convenient at times, to refer to these arrangements of operations as modules, without loss of generality. The described operations and their associated modules may be embodied in software, firmware, hardware, or any combinations thereof.

[00116] Any of the steps, operations, or processes described herein may be performed or implemented with one or more hardware or software modules, alone or in combination with other devices. In one embodiment, a software module is implemented with a computer program product comprising a computer-readable medium containing computer program code, which can be executed by a computer processor for performing any or all of the steps, operations, or processes described.

[00117] Embodiments of the invention may also relate to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, and/or it may comprise a general-purpose computing device selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a non-transitory, tangible computer readable storage medium, or any type of media suitable for storing electronic instructions, which may be coupled to a computer system bus. Furthermore,

any computing systems referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[00118] Embodiments of the invention may also relate to a product that is produced by a computing process described herein. Such a product may comprise information resulting from a computing process, where the information is stored on a non-transitory, tangible computer readable storage medium and may include any embodiment of a computer program product or other data combination described herein.

Finally, the language used in the specification has been principally selected for readability and instructional purposes, and it may not have been selected to delineate or circumscribe the inventive subject matter. It is therefore intended that the scope of the invention be limited not by this detailed description, but rather by any claims that issue on an application based hereon. Accordingly, the disclosure of the embodiments of the invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.

What is claimed is:

1. A method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, comprising:

receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders;

scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine;

generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests;

receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens;

scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and

generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request.

2. The method of claim 1, further comprising:

responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request.

3. The method of claim 2, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine.

4. The method of claim 1, wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request.

5. The method of claim 4, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request.

6. The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request.

7. The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests.

8. The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU).

9. The method of claim 1, wherein each token in the sequence of input tokens represents a text unit.

10. A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising:

receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders;

scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine;

generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests;

receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens;

scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and

generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request.

11. The non-transitory computer-readable storage medium of claim 10, the operations further comprising:

responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and

wherein the second batch of requests includes the remaining requests from the one or more requests and the new request.

12. The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine.

13. The non-transitory computer-readable storage medium of claim 10, wherein the second set of inputs include the sequence of input tokens for the new request and at least one output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request.

14. The non-transitory computer-readable storage medium of claim 13,
wherein the execution engine includes a cache memory for maintaining a key cache
tensor for storing keys and a value cache tensor for storing values for the at least
one request, and
wherein after scheduling the second batch of requests, allocating, by the execution
engine, a new cache memory dedicated to maintaining a key cache tensor and a
value cache tensor for the new request.

15. The non-transitory computer-readable storage medium of claim 14, wherein after
generating the second set of output tokens, a length of the key cache tensor for the at least one
request is different from a length of a key cache tensor for the new request, and a length of the
value cache tensor for the at least one request is different from a length of a value cache tensor
for the new request.

16. The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests.

17. The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU).

18. The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

ABSTRACT OF THE DISCLOSURE

An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length.

Electronic Patent Application Fee Transmittal				
Application Number:				
Filing Date:				
Title of Invention:		Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
First Named Inventor/Applicant Name:		Gyeongin Yu		
Filer:		Jae Yeon Baek/Carol Bergeron		
Attorney Docket Number:		39192-50678/US		
Filed as Small Entity				
Filing Fees for Track I Prioritized Examination - Nonprovisional Application under 35 USC 111(a)				
Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Basic Filing:				
UTILITY FILING FEE (ELECTRONIC FILING)	4011	1	80	80
UTILITY SEARCH FEE	2111	1	350	350
UTILITY EXAMINATION FEE	2311	1	400	400
REQUEST FOR PRIORITIZED EXAMINATION	2817	1	2100	2100
Pages:				
Claims:				
Miscellaneous-Filing:				
PROCESSING FEE, EXCEPT PROV. APPLS.	2830	1	70	70

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
Total in USD (\$)				3000

Electronic Acknowledgement Receipt

EFS ID:	44434105
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Yeon Baek
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	03-DEC-2021
Filing Date:	
Time Stamp:	18:29:33
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	\$3000
RAM confirmation Number	E2021B3I29563647
Deposit Account	192555
Authorized User	Jae Yeon Baek

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

37 CFR 1.19 (Document supply fees)

37 CFR 1.20 (Post Issuance fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	TrackOne Request	50678_Certification_and_Req uest_for_Prioritized_Examinatio n.pdf	114571	no	2
			a9901068d4a4917a3e6eaab736bd5c3973baf2b5		
Warnings:					
Information:					
2	Internet Communications Authorized	50678_Authorization_for_Inter net_Communications.pdf	75278	no	2
			565c10a43b4a5e8177d00221e861653a2a514700		
Warnings:					
Information:					
3	Application Data Sheet	50678_Application_Data_Shee t.pdf	1256148	no	9
			1d085613c8e673b05499030f14139a719a0025f9		
Warnings:					
Information:					
4	Drawings-only black and white line drawings	50678_Drawings.pdf	162285	no	12
			5c22e4b6d38d4006f1df037f37c64b08b8b7cfe3		
Warnings:					
Information:					
5		50678_Specification.pdf	262296	yes	57
			12cb27dc460b6b124e39c069dc7bb402c6a3efe4		
	Multipart Description/PDF files in .zip description				
	Document Description		Start	End	
	Specification		1	50	
	Claims		51	56	
	Abstract		57	57	
90 of 265					

Warnings:					
Information:					
6	Fee Worksheet (SB06)	fee-info.pdf	48677	no	2
			d2104116e6ac6e56c9fd445581b1200edb547b0d		
Warnings:					
Information:					
Total Files Size (in bytes):				1919255	
This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503. <u>New Applications Under 35 U.S.C. 111</u> If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application. <u>National Stage of an International Application under 35 U.S.C. 371</u> If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course. <u>New International Application Filed with the USPTO as a Receiving Office</u> If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.					

TRANSMITTAL FOR POWER OF ATTORNEY TO ONE OR MORE REGISTERED PRACTITIONERS

NOTE: This form is to be submitted with the Power of Attorney by Applicant form (PTO/AIA/82B) to identify the application to which the Power of Attorney is directed, in accordance with 37 CFR 1.5, unless the application number and filing date are identified in the Power of Attorney by Applicant form. If neither form PTO/AIA/82A nor form PTO/AIA/82B identifies the application to which the Power of Attorney is directed, the Power of Attorney will not be recognized in the application.

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Title	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
Art Unit	Not Yet Known
Examiner Name	Not Yet Known
Attorney Docket Number	39192-50678/US

SIGNATURE of Applicant or Patent Practitioner

Signature	/Jae Won Song/	Date (Optional)	December 7, 2021
Name	Jae Won Song	Registration Number	59,070
Title (if Applicant is a juristic entity)	Attorney for Applicant		
Applicant Name (if Applicant is a juristic entity)	FriendliAI Inc.		

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. If more than one applicant, use multiple forms.



*Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☒ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above:

00758

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☒ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name

Address

City

State

Zip

Country

Telephone

Email

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

FriendliAI Inc.

☐ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☒ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature

Date (Optional)

Name

Byung-Gon Chun

Title

CEO

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☒ Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

TRANSMITTAL FOR POWER OF ATTORNEY TO ONE OR MORE REGISTERED PRACTITIONERS

NOTE: This form is to be submitted with the Power of Attorney by Applicant form (PTO/AIA/82B) to identify the application to which the Power of Attorney is directed, in accordance with 37 CFR 1.5, unless the application number and filing date are identified in the Power of Attorney by Applicant form. If neither form PTO/AIA/82A nor form PTO/AIA82B identifies the application to which the Power of Attorney is directed, the Power of Attorney will not be recognized in the application.

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Title	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
Art Unit	Not Yet Known
Examiner Name	Not Yet Known
Attorney Docket Number	39192-50678/US

SIGNATURE of Applicant or Patent Practitioner

Signature	/Jae Won Song/	Date (Optional)	December 7, 2021
Name	Jae Won Song	Registration Number	59,070
Title (if Applicant is a juristic entity)	Attorney for Applicant		
Applicant Name (if Applicant is a juristic entity)	Seoul National University R&DB Foundation		

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. If more than one applicant, use multiple forms.



*Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☒ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above:

00758

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☒ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or
Individual Name

Address

City

State

Zip

Country

Telephone

Email

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

Seoul National University R&DB Foundation

- ☐ Inventor or Joint Inventor (title not required below)
- ☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)
- ☒ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)
- ☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature

Date (Optional)

2021. 12. 7

Name

Title

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☒ Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	44457324
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Won Song
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	07-DEC-2021
Filing Date:	
Time Stamp:	21:23:00
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no
------------------------	----

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Oath or Declaration filed	50678_Declaration.pdf	6824967	no	5
			eb5a4b461315a2a5e5bbfbee4bf125b7494bdf3c		

Warnings:

96 of 265

Information:					
2	Power of Attorney	50678_Power_of_Attorney_for _FriendlyAI.pdf	1693313	no	2
			3e2c4da72ce60a067191614494fc0a108939c647		
Warnings:					
Information:					
3	Power of Attorney	50678_Power_of_Attorney_for _SNU.pdf	1990927	no	2
			d7b8b08c4ef7bab9918e029767216810d784777a		
Warnings:					
Information:					
Total Files Size (in bytes):			10509207		
This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503. <u>New Applications Under 35 U.S.C. 111</u> If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application. <u>National Stage of an International Application under 35 U.S.C. 371</u> If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course. <u>New International Application Filed with the USPTO as a Receiving Office</u> If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.					

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of
Invention

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

As the below named inventor, I hereby declare that:

This declaration
is directed to:

☐

The attached application, or

☒

United States application or PCT international application number 17/542,193

filed on December 3, 2021

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Gyeongin Yu

Date (Optional) : _____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-8199 and select option 2.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of
Invention

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

As the below named inventor, I hereby declare that:

This declaration
is directed to:

☐

The attached application, or

☒

United States application or PCT international application number 17/542,193

filed on December 3, 2021

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

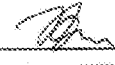
WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Geon-Woo Kim

Date (Optional): _____

Signature: 

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9198 and select option 2.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
--------------------	--

As the below named inventor, I hereby declare that:

This declaration is directed to: ☐ The attached application, or ☒ United States application or PCT international application number 17/542,193 filed on December 3, 2021

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

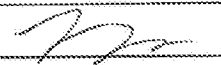
I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Joo Seong Jeong Date (Optional): _____

Signature: 

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9198 and select option 2.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of
Invention

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

As the below named inventor, I hereby declare that:

This declaration
is directed to:

☐

The attached application, or

☒

United States application or PCT international application number 17/542,193

filed on December 3, 2021

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

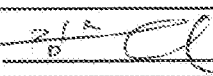
WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Soojeong Kim

Date (Optional) : _____

Signature: 

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9198 and select option 2.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of
Invention

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

As the below named inventor, I hereby declare that:

This declaration
is directed to:



The attached application, or



United States application or PCT international application number 17/542,193

filed on December 3, 2021

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Byung-Gon Chun

Date (Optional) : _____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875						Application or Docket Number 17/542,193				
APPLICATION AS FILED - PART I										
(Column 1)		(Column 2)		SMALL ENTITY		OR OTHER THAN SMALL ENTITY				
FOR	NUMBER FILED	NUMBER EXTRA	RATE(\$)	FEE(\$)		RATE(\$)	FEE(\$)			
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	80		N/A				
SEARCH FEE (37 CFR 1.16(k), (l), or (m))	N/A	N/A	N/A	350		N/A				
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	400		N/A				
TOTAL CLAIMS (37 CFR 1.16(j))	18	minus 20 = *	x 50 =	0.00	OR					
INDEPENDENT CLAIMS (37 CFR 1.16(h))	2	minus 3 = *	x 240 =	0.00						
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is \$310 (\$155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).			0.00						
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				0.00						
			TOTAL	830		TOTAL				
* If the difference in column 1 is less than zero, enter "0" in column 2.										
APPLICATION AS AMENDED - PART II										
(Column 1)		(Column 2)		(Column 3)		SMALL ENTITY		OR OTHER THAN SMALL ENTITY		
AMENDMENT A		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE(\$)	ADDITIONAL FEE(\$)		RATE(\$)	ADDITIONAL FEE(\$)
	Total (37 CFR 1.16(i))	*	Minus	**	=	x	=	OR	x	=
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x	=	OR	x	=
	Application Size Fee (37 CFR 1.16(s))							OR		
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	
(Column 1)		(Column 2)		(Column 3)		SMALL ENTITY		OR OTHER THAN SMALL ENTITY		
AMENDMENT B		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE(\$)	ADDITIONAL FEE(\$)		RATE(\$)	ADDITIONAL FEE(\$)
	Total (37 CFR 1.16(i))	*	Minus	**	=	x	=	OR	x	=
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x	=	OR	x	=
	Application Size Fee (37 CFR 1.16(s))							OR		
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	
* If the entry in column 1 is less than the entry in column 2, write "0" in column 3. ** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20". *** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3". The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.										



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY. DOCKET NO	TOT CLAIMS	IND CLAIMS
17/542,193	12/03/2021		910	39192-50678/US	18	2

CONFIRMATION NO. 1026

FILING RECEIPT



0000000130488086

758
FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

Date Mailed: 12/16/2021

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Gyeongin Yu, Seoul, KOREA, REPUBLIC OF;
Geon-Woo Kim, Seoul, KOREA, REPUBLIC OF;
Joo Seong Jeong, Seoul, KOREA, REPUBLIC OF;
Soojeong Kim, Seoul, KOREA, REPUBLIC OF;
Byung-Gon Chun, Seoul, KOREA, REPUBLIC OF;

Applicant(s)

FriendlyAI Inc., Seoul, KOREA, REPUBLIC OF;
Seoul National University R&DB Foundation, Seoul, KOREA, REPUBLIC OF;

Power of Attorney: The patent practitioners associated with Customer Number 00758

Domestic Applications for which benefit is claimed - None.

A proper domestic benefit claim must be provided in an Application Data Sheet in order to constitute a claim for domestic benefit. See 37 CFR 1.76 and 1.78.

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see <http://www.uspto.gov> for more information.) - None.

Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 12/14/2021

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 17/542,193**

Projected Publication Date: 06/08/2023

Non-Publication Request: No

Early Publication Request: No

**** SMALL ENTITY ****

Title

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

Preliminary Class

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at <http://www.uspto.gov/web/offices/pac/doc/general/index.html>.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, <http://www.stopfakes.gov>. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific

countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER
Title 35, United States Code, Section 184
Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop

technology, manufacture products, deliver services, and grow your business, visit <http://www.SelectUSA.gov> or call +1-202-482-6800.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026
758	7590	01/28/2022	EXAMINER	
FENWICK & WEST LLP			CENTRAL, DOCKET	
SILICON VALLEY CENTER			ART UNIT	
801 CALIFORNIA STREET			PAPER NUMBER	
MOUNTAIN VIEW, CA 94041			OPAP	
			NOTIFICATION DATE	DELIVERY MODE
			01/28/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

PTOC@Fenwick.com

<i>Decision Granting Request for Prioritized Examination (Track I)</i>	Application No. 17/542,193	Applicant(s) Yu et al.	
	Examiner APRIL M WISE	Art Unit OPET	AIA (FITF) Status Yes
1. THE REQUEST FILED <u>03 December 2021</u> IS GRANTED . The above-identified application has met the requirements for prioritized examination A. ☒ for an original nonprovisional application (Track I). B. ☐ for an application undergoing continued examination (RCE). 2. The above-identified application will undergo prioritized examination. The application will be accorded special status throughout its entire course of prosecution until one of the following occurs: A. filing a <u>petition for extension of time</u> to extend the time period for filing a reply; B. filing an <u>amendment to amend the application to contain more than four independent claims, more than thirty total claims</u>, or a multiple dependent claim; C. filing a <u>request for continued examination</u> ; D. filing a notice of appeal; E. filing a request for suspension of action; F. mailing of a notice of allowance; G. mailing of a final Office action; H. completion of examination as defined in 37 CFR 41.102; or I. abandonment of the application. Telephone inquiries with regard to this decision should be directed to undersigned at (571)272-1642. In his/her absence, calls may be directed to Petition Help Desk at (571) 272-3282.			
/APRIL M WISE/ Paralegal Specialist, Office of Petitions			

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Complete if Known

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Applicant	FriendliAI Inc.
Art Unit	2199
Confirmation Number	1026
Examiner Name	TANG, KENNETH
Attorney Docket Number	39192-50678/US

Sheet 1 of 3

U.S. PATENT DOCUMENTS

Examiner Initials*	Cite No. ¹	Document No. Number – Kind Code ² (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document

FOREIGN PATENT DOCUMENTS

Examiner Initials*	Cite No. ¹	Foreign Patent Document Country Code ³ – Number ⁴ Kind Code ⁵ (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	T ⁶

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published	T ⁶
	C1	FANG, J. et al., "TurboTransformers: An Efficient GPU Serving System For Transformer Models," arXiv:2010.05680v4, February 20, 2021, pp. 1-14.	
	C2	GAO, P. et al., "Low Latency RNN Inference with Cellular Batching," EuroSys '18, April 2018, pp. 1-15.	
	C3	GITHUB, "microsoft/DeepSpeed," January 19, 2021, pp. 1-9, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/microsoft/DeepSpeed >.	
	C4	GITHUB, "NVIDIA/FasterTransformer," April 2, 2021, pp. 1-28, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/NVIDIA/FasterTransformer >.	
	C5	GITHUB, "NVIDIA/Megatron-LM," August 11, 2021, pp. 1-18, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/NVIDIA/Megatron-LM >.	
	C6	LI, G. et al., "Easy and Efficient Transformer: Scalable Inference Solution for Large NLP Model," arXiv:2104.12470v4, November 23, 2021, pp. 1-9.	
	C7	NVIDIA, "NVIDIA TensorRT," January 27, 2021, pp. 1-11, [Online] [Retrieved on January 31, 2022] Retrieved from the Wayback Machine <URL: http://web.archive.org/web/20210127111124/https://developer.nvidia.com/tensorrt >.	

EXAMINER SIGNATURE

Examiner Signature	
Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.

⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Complete if Known

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Applicant	FriendliAI Inc.
Art Unit	2199
Confirmation Number	1026
Examiner Name	TANG, KENNETH
Attorney Docket Number	39192-50678/US

Sheet

2

of

3

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published	T ⁶
	C8	NVIDIA, "NVIDIA Triton Inference Server," January 25, 2021, pp. 1-6, [Online] [Retrieved on January 31, 2022] Retrieved from the Wayback Machine <URL: http://web.archive.org/web/20210125141031/https://developer.nvidia.com/nvidia-triton-inference-server >.	
	C9	OLSTON, C. et al., "TensorFlow-Serving: Flexible, High-Performance ML Serving," arXiv:1712.06139v2, December 27, 2017, pp. 1-8.	
	C10	SHAZEER, N. et al., "Mesh-TensorFlow: Deep Learning for Supercomputers," arXiv:1811.02084v1, November 5, 2018, pp. 1-16.	
	C11	SHOEYBI, M. et al., "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," arXiv:1909.08053v4, March 13, 2020, pp. 1-15.	
	C12	WANG, X. et al., "LightSeq: A High Performance Inference Library for Transformers," arXiv:2010.13887v4, April 22, 2021, pp. 1-8.	

EXAMINER SIGNATURE

Examiner Signature	
Date Considered	
*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.	
¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.	
⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.	

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Complete if Known

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Applicant	FriendliAI Inc.
Art Unit	2199
Confirmation Number	1026
Examiner Name	TANG, KENNETH
Attorney Docket Number	39192-50678/US

Sheet	3	of	3
-------	---	----	---

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ This application relies, under 35 U.S.C. § 120, on the earlier filing date of prior application No. _____, filed _____ and the references cited therein are hereby referenced, but are not required to be provided in this application under 37 CFR § 1.98(d).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Jae Won Song/	Date (YYYY-MM-DD)	2022-02-01
Name/Print	Jae Won Song	Registration Number	59,070

39192/50678/FW/15501046.1

Electronic Acknowledgement Receipt

EFS ID:	44886106
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Won Song/Christine Labella
Filer Authorized By:	Jae Won Song
Attorney Docket Number:	39192-50678/US
Receipt Date:	01-FEB-2022
Filing Date:	03-DEC-2021
Time Stamp:	10:17:35
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no
------------------------	----

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Information Disclosure Statement (IDS) Form (SB08)	50678_IDS.pdf	237920	no	3
			690e833c7e351baa1c3640305ab632f689c2c96a		

Warnings:

113 of 265

Information:					
This is not an USPTO supplied IDS fillable form					
2	Non Patent Literature	50677_Fang.pdf	12557014	no	14
			dc6f7bbe3c3f7741946925173ab3431d5e32d6fc		
Warnings:					
Information:					
3	Non Patent Literature	50677_Gao.pdf	14755631	no	15
			907361acb15ff0becb59115389f7f121280c18da		
Warnings:					
Information:					
4	Non Patent Literature	50677_Github_microsoft.pdf	731161	no	9
			027ef692694f92abf9a181dd20e132ecbc5977d7		
Warnings:					
Information:					
5	Non Patent Literature	50677_Github_Nvidia_Faster.pdf	2824505	no	28
			a5f79cb6c4117dc09348a8ef307af26633ff97ab		
Warnings:					
Information:					
6	Non Patent Literature	50677_Github_Nvidia_Megatron.pdf	1526596	no	18
			41afb8d8ce5ac9118b0ab6aa31564f0fe28425231		
Warnings:					
Information:					
7	Non Patent Literature	50677_Li.pdf	6889151	no	9
			38af685e7365f28133564177fb225f395a848831		
Warnings:					
Information:					
8	Non Patent Literature	50677_Nvidia_TensorRT.pdf	1737048	no	11
			a9698218bd5ba4b9816d61c165c699f7e3edb5f1		
Warnings:					
Information:					
114 of 265					

9	Non Patent Literature	50677_Nvidia_Triton.pdf	849793	no	6
			a82b8ec2492e5794c9ae9a8cd8403bba68ed22e9		
Warnings:					
Information:					
10	Non Patent Literature	50677_Olston.pdf	8905541	no	8
			dfaa3f4c592d8f1ae431bb9da4b6d100c3cac427		
Warnings:					
Information:					
11	Non Patent Literature	50677_Shazeer.pdf	11412381	no	16
			c5d41f17295e627b45a59bb306e5a52a09fa08ac		
Warnings:					
Information:					
12	Non Patent Literature	50677_Shoeysi.pdf	15284036	no	15
			5e3ed3ec97d3207ea9662f5f035acdf8d72af5f		
Warnings:					
Information:					
13	Non Patent Literature	50677_Wang.pdf	6216775	no	8
			701f2f2d988431a666d7cda470418cbb309cd76		
Warnings:					
Information:					
Total Files Size (in bytes):			83927552		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026
758	7590	02/09/2022	EXAMINER	
FENWICK & WEST LLP			TANG, KENNETH	
SILICON VALLEY CENTER			ART UNIT	
801 CALIFORNIA STREET			PAPER NUMBER	
MOUNTAIN VIEW, CA 94041			2199	
			NOTIFICATION DATE	DELIVERY MODE
			02/09/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

PTOC@Fenwick.com

Office Action Summary	Application No. 17/542,193	Applicant(s) Yu et al.	
	Examiner KENNETH TANG	Art Unit 2199	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 12/7/21.
☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on ____.
- 2a) ☐ This action is **FINAL**. 2b) ☒ This action is non-final.
- 3) ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on ____; the restriction requirement and election have been incorporated into this action.
- 4) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

- 5) ☒ Claim(s) 1-18 is/are pending in the application.
5a) Of the above claim(s) ____ is/are withdrawn from consideration.
- 6) ☐ Claim(s) ____ is/are allowed.
- 7) ☒ Claim(s) 1-4, 7-13 and 16-18 is/are rejected.
- 8) ☒ Claim(s) 5-6 and 14-15 is/are objected to.
- 9) ☐ Claim(s) ____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to PPHfeedback@uspto.gov.

Application Papers

- 10) ☐ The specification is objected to by the Examiner.
- 11) ☒ The drawing(s) filed on 12/3/21 is/are: a) ☒ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

- a) ☐ All b) ☐ Some** c) ☐ None of the:
1. ☐ Certified copies of the priority documents have been received.
2. ☐ Certified copies of the priority documents have been received in Application No. ____.
3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- 1) ☒ Notice of References Cited (PTO-892)
- 2) ☒ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b)
Paper No(s)/Mail Date 2/1/22.
- 3) ☐ Interview Summary (PTO-413)
Paper No(s)/Mail Date ____.
- 4) ☐ Other: ____.

DETAILED ACTION

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.
2. Claims 1-18 are presented for examination.

Specification

3. The abstract of the disclosure is objected to because:
 - In the Abstract, line 2, “internal sate” should be replaced with – internal state – to correct the minor typographical error.

Correction is required. See MPEP § 608.01(b).

Claim Objections

4. Claims 8 and 17 are objected to because of the following informalities: “tenor” should be replaced with – tensor -. Appropriate correction is required.

Claim Rejections - 35 USC § 103

5. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

6. Claims 1-4, 7-13, and 16-18 are rejected under 35 U.S.C. 103 as being unpatentable over Aarts et al. (hereinafter Aarts) (US 2021/0192314 A1) in view of Chung et al. (hereinafter Chung) (US 10,846,096 B1).

7. As to claim 1, **Aarts** teaches a method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model (*dynamic batch execution using machine learning model and training*) ([0361]; [0395]; [0545]), comprising:

receiving, by a serving system, one or more requests for execution, the serving system including a scheduler (*scheduler or pipeline manager 3912*) and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders (*machine learning systems and application including means for real-time language translation*) ([0489]; [0565]);

scheduling, by the scheduler, a batch (*Batch 606*) of requests including the one or more requests for execution on an execution engine ([0565]; *Fig. 6*);

generating, by the execution engine, a first set of output tokens by applying the transformer model (*trained machine learning model may be referred to as an output model 3816 that outputs in the deployment*) to a first set of inputs for the batch of requests (*input tensor 602 with different/variable lengths of sequence 608 relating to b1, b2, b3, b4 of batch 606*), wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests (*Fig. 6; [0541]; [0594]*);

receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens (*new training data or new input data coming in to the initial model 4204, etc.*) [0566]; [0594]);

scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch (*batch number is greater than 1, etc.*) of requests scheduled responsive to execute the second batch of requests ([0093]-[0094]; Fig. 6; [0541]; [0594]); and

generating, by the execution engine, a second set of output tokens by applying the transformer model (*trained machine learning model*) to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request (Fig. 6; [0541]; [0594]).

8. **Aarts** does not explicitly disclose its execution of a second batch based on memory available. However, **Chung** teaches batch processing of requests using a trained machine learning model and a resource manager to determine whether there are adequate resources available to suit the request, and if so, can provision the resources or otherwise grant access to the corresponding portion of those resources for use for an amount specified by the request. Furthermore, once a request is completed or pulled from a queue, the next request for the relevant service or model will be the first new request for the next or second batch (*col. 3, lines 41-46; col. 7, lines 42-51*). **Aarts** and **Chung** are analogous art with the claimed invention because they are all in the same field of endeavor batch processing using a machine learning model. It would have been obvious to one of ordinary skill in the art before the effective date of the application to modify **Aarts's** batch processing system/method such that it would include the feature of executing a second batch if there is memory available, as taught and suggested in **Chung's** batch processing system/method. The suggestion/motivation for doing so would have been to provide the predicted result of having accurate, near real-time services without significant

degradation in performance or increase in resource requirements by providing for the batch requests to be processed effectively using a learning model (*Chung – col. 3, lines 19-46*).

9. As to claim 2, **Aarts** ([0594]) and **Chung** (*col. 3, lines 41-46; col. 7, lines 42-51*) teaches further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request.

10. As to claim 3, **Aarts** teaches wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine ([0190]; [0293]; [0320]-[0321]; [0324]).

11. As to claim 4, **Aarts** teaches wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request (*variable sequence lengths for a given batch*) ([0093]; [0064]; [0406]).

12. As to claim 7, **Chung** teaches after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the

Art Unit: 2199

second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests (*col. 3, lines 41-46; col. 7, lines 42-51*).

13. As to claim 8, **Aarts** teaches wherein the execution engine is configured as a graphics processing unit (GPU) (*GPU(s) 1208*) or a tensor processing unit (TPU) (*tensor processing unit, TPC*) ([0174]; [0280]).

14. As to claim 9, **Aarts** teaches wherein each token in the sequence of input tokens represents a text unit (*sequences for text string comparison*) ([0406]; [0096]).

15. As to claim 10, it is rejected for the same reasons as stated in the rejection of claim 1.

16. As to claim 11, it is rejected for the same reasons as stated in the rejection of claim 2.

17. As to claim 12, it is rejected for the same reasons as stated in the rejection of claim 3.

18. As to claim 13, it is rejected for the same reasons as stated in the rejection of claim 4.

19. As to claim 16, it is rejected for the same reasons as stated in the rejection of claim 7.

Art Unit: 2199

20. As to claim 17, it is rejected for the same reasons as stated in the rejection of claim 8.

21. As to claim 18, it is rejected for the same reasons as stated in the rejection of claim 9.

Allowable Subject Matter

22. Claims 5-6 and 14-15 are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.

Conclusion

23. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure:

- **Choi et al. ("PREMA: A Predictive Multi-task Scheduling Algorithm For Preemptible Neural Processing Units", 2020)** teaches a deep neural network (DNN) based machine learning algorithms with variable sequence lengths of data.

24. Any inquiry concerning this communication or earlier communications from the examiner should be directed to KENNETH TANG whose telephone number is (571)272-3772. The examiner can normally be reached Monday-Friday 7AM-3PM.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at <http://www.uspto.gov/interviewpractice>.

Art Unit: 2199

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Lewis Bullock can be reached on 571-272-3759. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit:

<https://patentcenter.uspto.gov>. Visit <https://www.uspto.gov/patents/apply/patent-center> for more information about Patent Center and <https://www.uspto.gov/patents/docx> for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/KENNETH TANG/

Primary Examiner, Art Unit 2199

<i>Notice of References Cited</i>	Application/Control No. 17/542,193		Applicant(s)/Patent Under Reexamination Yu et al.	
	Examiner KENNETH TANG		Art Unit 2199	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20210192314-A1	06-2021	Aarts; Bastiaan Joannes Matheus	G06F8/433	1/1
*	B	US-10846096-B1	11-2020	Chung; Kiuk	G06N20/00	1/1
*	C	US-20210406673-A1	12-2021	Pardeshi; Siddhant	G06N3/08	1/1
*	D	US-20200311341-A1	10-2020	Chaturvedi; Snigdha	G06N20/00	1/1
*	E	US-20210279576-A1	09-2021	Shazeer; Noam M.	G06N3/08	1/1
*	F	US-20210263779-A1	08-2021	Haghighat; Mohammad R.	G06F11/3409	1/1
*	G	US-20200226453-A1	07-2020	Luk; Eric	G06N3/08	1/1
	H					
	I					
	J					
	K					
	L					
	M					

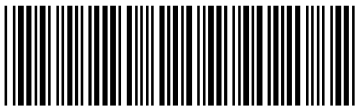
FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Choi et al., "PREMA: A Predictive Multi-task Scheduling Algorithm For Preemptible Neural Processing Units", 2020, IEEE International Symposium on High Performance Computer Architecture (HPCA), pgs. 220-233. (Year: 2020)
	V	Dai et al, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context", 6/2/19, Carnegie Mellon, University, Google Brain, pgs. 1-20 (Year: 2019)
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

<i>Search Notes</i> 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

CPC - Searched*		
Symbol	Date	Examiner
G06F9/4881, G06N20/00, G06F9/5016, G06N5/04, G06N3/0454, G06N3/08	02/01/2022	KT

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner
718	1	02/01/2022	KT

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
PE2E search	02/01/2022	KT
Inventor search	02/01/2022	KT
IP.com search	02/01/2022	KT
Google Patents and NPL Scholar search	02/01/2022	KT

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner

/KENNETH TANG/ Primary Examiner, Art Unit 2199	
---	--

Bibliographic Data

Application No: 17/542,193

Foreign Priority claimed: ☐ Yes ☒ No

35 USC 119 (a-d) conditions met: ☐ Yes ☐ No ☐ Met After Allowance

Verified and Acknowledged: /KENNETH TANG/

Examiner's Signature

Initials

Title:

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
12/03/2021	718	2199	39192-50678/US
RULE			

APPLICANTS

FriendlyAI Inc., Seoul, KOREA, REPUBLIC OF

Seoul National University R&DB Foundation, Seoul, KOREA, REPUBLIC OF

INVENTORS

Gyeongin Yu, Seoul, KOREA, REPUBLIC OF

Geon-Woo Kim, Seoul, KOREA, REPUBLIC OF

Joo Seong Jeong, Seoul, KOREA, REPUBLIC OF

Soojeong Kim, Seoul, KOREA, REPUBLIC OF

Byung-Gon Chun, Seoul, KOREA, REPUBLIC OF

CONTINUING DATA

FOREIGN APPLICATIONS

IF REQUIRED, FOREIGN LICENSE GRANTED**

12/14/2021

**** SMALL ENTITY ****

STATE OR COUNTRY

KOREA, REPUBLIC OF

ADDRESS

FENWICK & WEST LLP

SILICON VALLEY CENTER

801 CALIFORNIA STREET

MOUNTAIN VIEW, CA 94041

UNITED STATES

FILING FEE RECEIVED

\$3,000

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Complete if Known

Sheet 1 of 3				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
				Attorney Docket Number	39192-50678/US

U.S. PATENT DOCUMENTS

Examiner Initials*	Cite No. ¹	Document No.	Number – Kind Code ² (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document

FOREIGN PATENT DOCUMENTS

Examiner Initials*	Cite No. ¹	Foreign Patent Document	Country Code ³ – Number ⁴ Kind Code ⁵ (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	T ⁶

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published	T ⁶
	C1	FANG, J. et al., "TurboTransformers: An Efficient GPU Serving System For Transformer Models," arXiv:2010.05680v4, February 20, 2021, pp. 1-14.	
	C2	GAO, P. et al., "Low Latency RNN Inference with Cellular Batching," EuroSys '18, April 2018, pp. 1-15.	
	C3	GITHUB, "microsoft/DeepSpeed," January 19, 2021, pp. 1-9, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/microsoft/DeepSpeed >.	
	C4	GITHUB, "NVIDIA/FasterTransformer," April 2, 2021, pp. 1-28, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/NVIDIA/FasterTransformer >.	
	C5	GITHUB, "NVIDIA/Megatron-LM," August 11, 2021, pp. 1-18, [Online] [Retrieved on January 31, 2022] Retrieved from the Internet <URL: https://github.com/NVIDIA/Megatron-LM >.	
	C6	LI, G. et al., "Easy and Efficient Transformer: Scalable Inference Solution for Large NLP Model," arXiv:2104.12470v4, November 23, 2021, pp. 1-9.	
	C7	NVIDIA, "NVIDIA TensorRT," January 27, 2021, pp. 1-11, [Online] [Retrieved on January 31, 2022] Retrieved from the Wayback Machine <URL: http://web.archive.org/web/20210127111124/https://developer.nvidia.com/tensorrt >.	

EXAMINER SIGNATURE

Examiner Signature	/KENNETH TANG/
Date Considered	02/04/2022
*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.	
¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.	
⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	2	of	3	Attorney Docket Number	39192-50678/US

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS				
Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published	T ⁶	
	C8	NVIDIA, "NVIDIA Triton Inference Server," January 25, 2021, pp. 1-6, [Online] [Retrieved on January 31, 2022] Retrieved from the Wayback Machine <URL: http://web.archive.org/web/20210125141031/https://developer.nvidia.com/nvidia-triton-inference-server >.		
	C9	OLSTON, C. et al., "TensorFlow-Serving: Flexible, High-Performance ML Serving," arXiv:1712.06139v2, December 27, 2017, pp. 1-8.		
	C10	SHAZEER, N. et al., "Mesh-TensorFlow: Deep Learning for Supercomputers," arXiv:1811.02084v1, November 5, 2018, pp. 1-16.		
	C11	SHOEYBI, M. et al., "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," arXiv:1909.08053v4, March 13, 2020, pp. 1-15.		
	C12	WANG, X. et al., "LightSeq: A High Performance Inference Library for Transformers," arXiv:2010.13887v4, April 22, 2021, pp. 1-8.		

EXAMINER SIGNATURE	
Examiner Signature	
Date Considered	
*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.	
¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.	
⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendlyAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	3	of	3	Attorney Docket Number	39192-50678/US

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ This application relies, under 35 U.S.C. § 120, on the earlier filing date of prior application No. _____, filed _____ and the references cited therein are hereby referenced, but are not required to be provided in this application under 37 CFR § 1.98(d).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Jae Won Song/	Date (YYYY-MM-DD)	2022-02-01
Name/Print	Jae Won Song	Registration Number	59,070

39192/50678/FW/15501046.1

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L1	0	17/542193.app.	(US-PGPUB)	OR	ON	ON	2022/01/30 08:33 PM
L3	1	"20210192314".pn.	(US-PGPUB)	OR	ON	ON	2022/01/31 01:49 PM
L4	1	("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/01/31 01:55 PM
L5	1	"20210192314".pn. AND (token text)	(US-PGPUB)	OR	ON	ON	2022/01/31 03:31 PM
L6	1	"20210192314".pn. AND (token text sequence)	(US-PGPUB)	OR	ON	ON	2022/01/31 03:31 PM
L7	444545	(718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/01/31 04:49 PM
L8	6039	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/01/31 04:50 PM
L9	21	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4))	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL,	OR	ON	ON	2022/01/31 04:50 PM

L10	21	AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. ((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/01/31 04:51 PM
L11	5	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:51 PM
L12	5	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:52 PM
L13	4	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:52 PM

L14	0	AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/01/31 04:52 PM
L15	4	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key AND sequence	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:53 PM
L16	2	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key AND sequence	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:54 PM
L17	4800	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:57 PM
L18	2122	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:57 PM

L19	708	AND schedul\$4 ((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:57 PM
L20	656	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:57 PM
L21	232	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:57 PM
L22	40	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu) AND length	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:58 PM
L23	38	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu) AND (length near5 (tensor key cache request different))	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 05:02 PM

		(tensor key cache request different)) AND decoder					
L24	1	"20210192314".pn. AND (tpu gpu)	(US-PGPUB)	OR	ON	ON	2022/01/31 05:13 PM
L25	1	"20210192314".pn. AND length	(US-PGPUB)	OR	ON	ON	2022/01/31 05:16 PM
L26	1	"20210192314".pn. AND (key AND cache)	(US-PGPUB)	OR	ON	ON	2022/01/31 05:17 PM
L27	1	"20210192314".pn. AND (tensor WITH cache)	(US-PGPUB)	OR	ON	ON	2022/01/31 05:18 PM
L28	1	"20210192314".pn. AND (dynamic\$4 (batch\$4 WITH request) (learn\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 02:56 PM
L29	1	"20210192314".pn. AND batch\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 02:59 PM
L30	1	"20210192314".pn. AND decod\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 03:03 PM
L31	7	("20200285977" "20200285912" "20190220703" "20210240825" "20190138934" "20210240826" "20210185066").pn. AND ((variable different) near5 length)	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/02 03:40 PM
L32	1	"20210192314".pn. AND cache	(US-PGPUB)	OR	ON	ON	2022/02/02 04:11 PM
L33	1	"20210192314".pn. AND (cache SAME (key tensor length value request batch))	(US-PGPUB)	OR	ON	ON	2022/02/02 04:12 PM
L34	1	"20210192314".pn. AND (gpu tpc tcu)	(US-PGPUB)	OR	ON	ON	2022/02/02 04:15 PM
L35	1	"20210192314".pn. AND (token sequence text)	(US-PGPUB)	OR	ON	ON	2022/02/02 04:17 PM
L36	1	"20210192314".pn. AND insufficien\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 04:19 PM
L37	1	"20210192314".pn. AND availab\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 04:19 PM
L38	1	("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/02/02 04:23 PM
L39	1	"20210192314".pn. AND transform\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 05:11 PM
L40	1	"20210192314".pn. AND availab\$4	(US-PGPUB)	OR	ON	ON	2022/02/02 05:20 PM
L41	1	"20210192314".pn.	(US-PGPUB)	OR	ON	ON	2022/02/02

L42	0	AND weight "20210192314".pn. AND (decoder SAME (model learn\$4 ai))	(US-PGPUB)	OR	ON	ON	10:53 PM 2022/02/02 10:55 PM
L43	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai))	(US-PGPUB)	OR	ON	ON	2022/02/02 10:55 PM
L44	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 10:56 PM
L45	0	"20210192314".pn. AND (decoder SAME (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 10:56 PM
L46	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 10:56 PM
L47	1	"20210192314".pn. AND (schedul\$4 AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 11:06 PM
L48	1	"20210192314".pn. AND (schedul\$4 SAME (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/02/02 11:06 PM
L49	1	"20210192314".pn. AND (schedul\$4 WITH (batch request))	(US-PGPUB)	OR	ON	ON	2022/02/02 11:12 PM
L50	1	"20210192314".pn. AND batch	(US-PGPUB)	OR	ON	ON	2022/02/02 11:13 PM
L51	1	"20210192314".pn. AND (output WITH (token model transform\$4 translat\$4 learn\$4 batch request))	(US-PGPUB)	OR	ON	ON	2022/02/02 11:16 PM
L52	1	"20210192314".pn. AND length	(US-PGPUB)	OR	ON	ON	2022/02/03 03:45 PM
L53	1	"20210192314".pn. AND ((new second) WITH (request batch))	(US-PGPUB)	OR	ON	ON	2022/02/03 03:47 PM
L54	1	"20210192314".pn. AND (split\$4 concatenat\$4)	(US-PGPUB)	OR	ON	ON	2022/02/03 04:00 PM
L55	1	"20210192314".pn. AND cache	(US-PGPUB)	OR	ON	ON	2022/02/03 04:06 PM
L56	0	"20190251418".pn. AND ((memory resource) WITH (sufficent insufficient enough availab\$4))	(US-PGPUB)	OR	ON	ON	2022/02/03 04:34 PM
L57	0	"20190251418".pn.	(US-PGPUB)	OR	ON	ON	2022/02/03

		AND ((memory resource) SAME (sufficient insufficient enough availab\$4))					04:34 PM
L58	1	"20190251418".pn. AND ((sufficient insufficient enough availab\$4))	(US-PGPUB)	OR	ON	ON	2022/02/03 04:34 PM
L59	1	"20210192314".pn. AND remain\$4	(US-PGPUB)	OR	ON	ON	2022/02/03 04:35 PM
L60	1	"20210192314".pn. AND (second new)	(US-PGPUB)	OR	ON	ON	2022/02/03 04:37 PM
L61	146	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:02 PM
L62	10	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:02 PM
L63	10	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:02 PM
L64	9	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder AND @ad<"20211203" AND ((availab\$4 sufficien\$4 insufficien\$4 threshold) near4 (capacity memory resource))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:11 PM
L65	14	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn.	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:19 PM

L66	14	AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight sufficien\$4 insufficien\$4 threshold availab\$4 length split\$4 concaten\$4 cache) ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:34 PM
L67	0	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:34 PM
L68	6	AND G06F9/5016.pn. ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176"	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:35 PM

L69	6	"20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. AND decoder	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:36 PM
L70	3	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868"	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 03:46 PM

		"20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. AND decoder AND keys					
L71	4	("10846096" "20210334134" "20210232399" "9652287").pn. AND (new WITH (request batch))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 04:35 PM
L72	1	("10846096").pn. AND (new WITH (request batch)) AND (memory resource capacity) WITH (availab\$4 sufficien\$4 insufficien\$4 threshold)	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 04:38 PM
L73	1	("10846096").pn. AND (new WITH (request batch)) AND ((memory resource capacity) WITH (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 04:38 PM
L74	1	"20210192314".pn. AND new	(US-PGPUB)	OR	ON	ON	2022/02/04 04:48 PM
L75	1	"20210192314".pn. AND (availab\$4 sufficien\$4 insufficien\$4 threshold)	(US-PGPUB)	OR	ON	ON	2022/02/04 04:58 PM
L76	11534	((variable different) near4 length) AND L7	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR,	OR	ON	ON	2022/02/04 05:24 PM

L77	9301	((variable different) near4 length) AND L7 AND learn\$4	TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:24 PM
L78	835	((variable different) near4 length) AND L7 AND learn\$4 AND tensor	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:24 PM
L79	430	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:25 PM
L80	93	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL,	OR	ON	ON	2022/02/04 05:25 PM

L81	92	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203"	NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:25 PM
L82	1	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND (key near3 (cache tensor))	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:25 PM
L83	70	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND (decod\$4 transform\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:26 PM
L84	0	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND	(DERWENT)	OR	ON	ON	2022/02/04 05:26 PM

L85	410	(decod\$4 transform\$4) batch\$4 AND learn\$4 AND (decod\$4 transform\$4)	(DERWENT)	OR	ON	ON	2022/02/04 05:27 PM
L86	51	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7	(DERWENT)	OR	ON	ON	2022/02/04 05:27 PM
L87	51	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203"	(DERWENT)	OR	ON	ON	2022/02/04 05:27 PM
L88	23	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu tcu)	(DERWENT)	OR	ON	ON	2022/02/04 05:28 PM
L89	30	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu normaliz\$4 tcu)	(DERWENT)	OR	ON	ON	2022/02/04 05:28 PM
L90	4	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu normaliz\$4 tcu) AND batch.ti.	(DERWENT)	OR	ON	ON	2022/02/04 05:29 PM
L91	32	(Artificial Intelligence) More Like: US-20210216870-A1 Text: (CA-2975251-C OR US-10417562-B2 OR JP-2021192251-A OR JP-6653343-B2 OR US- 10902319-B2 OR US- 20200012942-A1 OR US-20210357756-A1 OR JP-6935542-B2 OR US-20200234127-A1 OR AU-2019200309-B2 OR US-20200057924- A1 OR US-10628710- B2 OR US- 20160217368-A1 OR KR-102204286-B1 OR JP-2020506488-A OR CN-107278310-A OR JP-6876814-B2 OR CN-108537326-A OR KR-102300614-B1 OR	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:30 PM

L92	1	EP-3251059-B1).did. ("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/02/04 05:30 PM
L93	110173	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:32 PM
L94	5802	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:32 PM
L95	5802	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:32 PM
L96	701	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND tensor	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT,	OR	ON	ON	2022/02/04 05:32 PM

L97	15	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND tensor AND (variable near4 length)	JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:33 PM
L98	1	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND tensor AND (variable near4 length)	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 05:33 PM
L99	0	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND (batch WITH variable WITH length WITH request)	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 05:33 PM
L100	3	(batch WITH variable WITH length WITH request)	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 05:33 PM
L101	25	(batch\$4 near3 request) AND (variable WITH length WITH (model learn\$4))	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 05:34 PM
L102	22	batch\$4 WITH dynamic\$4 WITH (inference nerual learn\$4) WITH (transform\$4 decod\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG,	OR	ON	ON	2022/02/04 05:35 PM

L103	22	L102 AND @ad<"20211203"	SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/02/04 05:36 PM
L104	1	L102 AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/02/04 05:36 PM

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

Web Search History

date, time	web site	search string
2/1/2022 11:39:14 AM	Google Patents	Keyword Group 1 [batch scheduler and learning]
2/1/2022 11:39:23 AM	Google Patents	Keyword Group 1 [batch]
2/1/2022 11:39:25 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler]
2/1/2022 11:39:32 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning]
2/1/2022 11:39:35 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"]
2/1/2022 11:39:41 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"]
2/1/2022 11:39:54 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length]

date, time	web site	search string
2/1/2022 11:40:00 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu]
2/1/2022 11:40:01 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu]
2/1/2022 11:40:14 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [toekn]
2/1/2022 11:40:16 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token]

date, time	web site	search string
2/1/2022 11:40:17 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text]
2/1/2022 11:40:44 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Before Priority Date: priority:20211203
2/1/2022 11:41:15 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Before Priority Date: priority:20211203

date, time	web site	search string
2/1/2022 11:41:34 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] Before Priority Date: priority:20211203
2/1/2022 11:41:40 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] Before Priority Date: priority:20211203

date, time	web site	search string
2/1/2022 11:41:52 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] OR [encoder] Before Priority Date: priority:20211203
2/1/2022 11:42:28 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] OR [encoder] Keyword Group 8 [memory] Before Priority Date: priority:20211203

date, time	web site	search string
2/1/2022 11:42:43 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] OR [encoder] Keyword Group 8 [memory] OR [cache] Before Priority Date: priority:20211203
2/1/2022 11:42:46 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] OR [encoder] Keyword Group 8 [memory] OR [cache] Patent Office(s) [US] Before Priority Date: priority:20211203

date, time	web site	search string
2/1/2022 11:43:05 AM	Google Patents	Keyword Group 1 [batch] Keyword Group 2 [scheduler] Keyword Group 3 [learning] OR ["deep learning"] OR ["artificial intelligence"] OR ["machine learning"] Keyword Group 4 [length] Keyword Group 5 [gpu] OR [tpu] Keyword Group 6 [token] OR [text] Keyword Group 7 [decoder] OR [translation] OR [encoder] Keyword Group 8 [memory] OR [cache] OR [insufficient] Patent Office(s) [US] Before Priority Date: priority:20211203
2/1/2022 11:46:20 AM	Google Scholar	batch scheduler learning,"deep learning","artificial intelligence","machine learning" length gpu,tpu token,text decoder,translation,encoder memory,cache,insufficient



Search History

Report Run Date: 2022-02-01 UTC

Queries: 15

Markers: 0

Query #	Timestamp	Results	Page
1	2022-01-31 17:49 UTC	15,676,685	3
2	2022-01-31 17:49 UTC	~15,675,107	5
3	2022-01-31 17:50 UTC	~820,689	5
4	2022-01-31 17:50 UTC	~93,595	5
5	2022-01-31 18:28 UTC	~90,546	5
6	2022-01-31 18:29 UTC	~87,133	6
7	2022-01-31 18:29 UTC	~38,553	6
8	2022-01-31 18:33 UTC	~38,555	8
9	2022-01-31 18:38 UTC	~18,128	9
10	2022-01-31 18:39 UTC	~4624	9
11	2022-01-31 18:41 UTC	~4624	10
12	2022-01-31 18:41 UTC	~2239	10
13	2022-01-31 22:02 UTC	~1902	10
14	2022-01-31 22:04 UTC	~2253	11
15	2022-01-31 22:10 UTC	~6225	12



US Patents | US Designs | US Applications | EPO Patents | EPO Applications | China Patents | China Applications | Japan Patents | Japan Applications | Korea Patents | Korea Applications | WIPO Applications | Argentina Patents | Argentina Applications | Brazil Patents | Brazil Applications | Canada Patents | Canada Applications | Chile Patents | Chile Applications | Colombia Applications | Costa Rica Applications | Cuba Patents | Cuba Applications | Dominican Republic Applications | Ecuador Patents | Ecuador Applications | El Salvador Applications | Guatemala Applications | Honduras Applications | Mexico Patents | Mexico Applications | Nicaragua Patents | Panama Applications | Peru Applications | Trinidad & Tobago Patents | Uruguay Applications | Austria Patents | Austria Applications | Belarus Patents | Belgium Patents | Belgium Applications | Bosnia & Herzegovina Patents | Bosnia & Herzegovina Applications | Bulgaria Patents | Bulgaria Applications | Croatia Patents | Croatia Applications | Czech Republic Patents | Czech Republic Applications | Czechoslovakia Patents | Czechoslovakia Applications | Denmark Patents | Denmark Applications | Estonia Patents | Estonia Applications | EUIPO Patents | Finland Patents | Finland Applications | France Patents | France Applications | Germany Patents | Germany Applications | Great Britain Patents | Great Britain Applications | Greece Patents | Greece Applications | Hungary Patents | Hungary Applications | Iceland Patents | Iceland Applications | Ireland Patents | Ireland Applications | Italy Patents | Italy Applications | Latvia Patents | Latvia Applications | Lithuania Patents | Lithuania Applications | Luxembourg Patents | Malta Patents | Monaco Patents | Montenegro Patents | Montenegro Applications | Netherlands Patents | Netherlands Applications | Norway Patents | Norway Applications | Poland Patents | Poland Applications | Portugal Patents | Portugal Applications | Republic of Moldova Patents | Republic of Moldova Applications | Romania Patents | Romania Applications | San Marino Patents | San Marino Applications | Serbia Patents | Serbia Applications | Slovakia Patents | Slovakia Applications | Slovenia Patents | Spain Patents | Spain Applications | Sweden Patents | Sweden Applications | Switzerland Patents | Switzerland Applications | Ukraine Patents | Yugoslavia/Serbia and Montenegro Patents | Yugoslavia/Serbia and Montenegro Applications | Armenia Patents | Australia Patents | Australia Applications | Cyprus Patents | Gulf Cooperation Council Patents | Hong Kong Patents | India Patents | India Applications | Indonesia Patents | Indonesia Applications | Israel Applications | Jordan Patents | Kyrgyzstan Patents | Macao Applications | Malaysia Patents | Mongolia Patents | New Zealand Patents | Philippines Patents | Philippines Applications | Saudi Arabia Patents | Singapore Patents | Singapore Applications | Taiwan Patents | Taiwan Applications | Tajikistan Patents | Tajikistan Applications | Thailand Applications | Uzbekistan Patents | Vietnam Patents | Algeria Patents | ARIPO Patents | ARIPO Applications | Egypt Patents | Kenya Patents | Malawi Patents | Morocco Patents | Morocco Applications | OAPI Patents | South Africa Patents | Tunisia Applications | Zambia Patents | Zimbabwe Patents | EAPO Patents | EAPO Applications | Georgia Patents | Georgia Applications | Kazakhstan Patents | Kazakhstan Applications | Russia Patents | Russia Applications | Turkey Patents | Turkey Applications

1 2022-01-31 17:49 UTC | 15,676,685 results from

De-Dup: None | Cut-off: None | Sort: Relevance

Main Concept: Text

1. A method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution

engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 2. The method of claim 1, further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 3. The method of claim 2, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 4. The method of claim 1, wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request. 5. The method of claim 4, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 6. The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 7. The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 8. The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 9. The method of claim 1, wherein each token in the sequence of input tokens represents a text unit. 10. A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 11. The non-transitory computer-readable storage medium of claim 10, the operations further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 12. The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 13. The non-transitory computer-readable storage medium of claim 10, wherein the second set of inputs include the sequence of input tokens for the new request and at least one output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request. 14. The non-transitory computer-readable storage medium of claim 13, wherein the execution engine includes a cache memory for

maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 15. The non-transitory computer-readable storage medium of claim 14, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 16. The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 17. The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 18. The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

Concept Modifiers: None

Filters: None

2 2022-01-31 17:49 UTC | ~15,675,107 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Query 1](#)

Concept Modifiers: None

Filters: None

3 2022-01-31 17:50 UTC | ~820,689 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Query 1](#)

Concept Modifiers (1):

[1. More Like Text:](#) machine learning

Filters: None

4 2022-01-31 17:50 UTC | ~93,595 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Query 1](#)

Concept Modifiers: [Same as Query 3](#)

Filters (1):

[Classic Syntax limited to:](#) encod* and decod*

5 2022-01-31 18:28 UTC | ~90,546 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Query 1](#)

Concept Modifiers (2):

2. [More Like Text](#): ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. [More Like Text](#): machine learning

Filters: [Same as Query 4](#)

6 2022-01-31 18:29 UTC | ~87,133 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Query 1](#)

Concept Modifiers: [Same as Query 5](#)

Filters (2):

[Classic Syntax limited to](#): encod* and decod*

[Earliest Priority Date](#): From 1997-05-01 To 2021-12-03

 Content: Patent Publications (2)

US Patents | US Applications

7 2022-01-31 18:29 UTC | ~38,553 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Text

1. A method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch

of requests including the sequence of input tokens for the new request. 2. The method of claim 1, further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 3. The method of claim 2, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 4. The method of claim 1, wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request. 5. The method of claim 4, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 6. The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 7. The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 8. The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 9. The method of claim 1, wherein each token in the sequence of input tokens represents a text unit. 10. A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 11. The non-transitory computer-readable storage medium of claim 10, the operations further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 12. The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 13. The non-transitory computer-readable storage medium of claim 10, wherein the second set of inputs include the sequence of input tokens for the new request and at least one output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request. 14. The non-transitory computer-readable storage medium of claim 13, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request,

and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 15. The non-transitory computer-readable storage medium of claim 14, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 16. The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 17. The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 18. The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

Concept Modifiers (2):

2. More Like Text: ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. More Like Text: machine learning

Filters (2):

Classic Syntax limited to: encod* and decod*

Earliest Priority Date: From 1997-05-01 To 2021-12-03

8 2022-01-31 18:33 UTC | ~38,555 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Same as Query 7

Concept Modifiers (3):

3. More Like Text: In many instances, the inputs for requests or the internal states for the requests are variable in length and may be difficult to batch using, for example, the batching method in FIGs. 2A-2B. For example, the number of input tokens between different NLP requests are highly likely to vary in length depending on the input text and the type of request. For example, the length of a first query request may be different from the length of a second query request since different users synthesize language differently. However, methods of batching transformer models, such as the method described in conjunction with FIGs. 2A-2B, require that the lengths of input tensors for the batch be the same since the input tensors have to be coalesced into a single concatenated tensor for an operation. Thus, if the lengths of input tensors are different for multiple requests, it may not be feasible to process the requests as a batch using the batching method of FIGs. 2A-2B. [0049] In addition, a workaround addressing this problem is to treat the input token sequences during the encoding phase as if the requests had the same lengths as the input token sequence with the shortest length (i.e., least number of tokens). However, this requires the inference system 130 to discard output tokens generated for requests with longer input lengths at subsequent iterations since the original input tokens received for the request have to be substituted in the respective elements. This results in inefficiency and waste of unnecessary resources because an inference system processes the input token sequences for longer requests

separately in multiple iterations instead of a single iteration. Another workaround may be to treat the input token sequences as if the requests had the same lengths as the input token sequence with the longest length (i.e., largest number of tokens). However, this requires the inference system to increase the size of the input tensors to match the request with the longest token sequence (and pad the remaining elements). This results in redundant computation since the inference system processes a larger input tensor throughout the process of executing the transformer model. For example, in some cases, the number of iterations required to process a batch of requests using the method of FIGs. 2A-2B may be significantly larger than the number of iterations to process each request individually. [0050] Thus, in one embodiment, the inference system 130 trains and applies a machine-learning transformer model to a batch of requests with variable input length or target length or internal state length using a selective batching method. In the selective batching method, the inference system 130 selectively batches a subset of operations in the transformer model but separately processes requests individually for a subset of operations in the transformer model that are not compatible with batching. In one embodiment, the operation to be processed separately is the attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system 130 can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing problems that occur in workarounds.

2. More Like Text: ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. More Like Text: machine learning

Filters: Same as Query 7

9 2022-01-31 18:38 UTC | ~18,128 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: Same as Query 7

Concept Modifiers: Same as Query 8

Filters (3):

Classic Syntax limited to: encod* and decod*

Earliest Priority Date: From 1997-05-01 To 2021-12-03

Classic Syntax limited to: schedul*

10 2022-01-31 18:39 UTC | ~4624 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: Same as Query 7

Concept Modifiers: Same as Query 8

Filters (4):

Classic Syntax limited to: encod* and decod*

Earliest Priority Date: From 1997-05-01 To 2021-12-03

Classic Syntax limited to: schedul*

Classic Syntax limited to: batch*

11 2022-01-31 18:41 UTC | ~4624 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Same as Query 7

Concept Modifiers: Same as Query 8

Filters: Same as Query 10

12 2022-01-31 18:41 UTC | ~2239 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Same as Query 7

Concept Modifiers: Same as Query 8

Filters (5):

Classic Syntax limited to: encod* and decod*

Earliest Priority Date: From 1997-05-01 To 2021-12-03

Classic Syntax limited to: schedul*

Classic Syntax limited to: batch*

Classic Syntax limited to: length

13 2022-01-31 22:02 UTC | ~1902 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Same as Query 7

Concept Modifiers (4):

4. More Like Text: decoder

3. More Like Text: In many instances, the inputs for requests or the internal states for the requests are variable in length and may be difficult to batch using, for example, the batching method in FIGs. 2A-2B. For example, the number of input tokens between different NLP requests are highly likely to vary in length depending on the input text and the type of request. For example, the length of a first query request may be different from the length of a second query request since different users synthesize language differently. However, methods of batching transformer models, such as the method described in conjunction with FIGs. 2A-2B, require that the lengths of input tensors for the batch be the same since the input tensors have to be coalesced into a single concatenated tensor for an operation. Thus, if the lengths of input tensors are different for multiple requests, it may not be feasible to process the requests as a batch using the batching method of FIGs. 2A-2B. [0049] In addition, a workaround addressing this problem is to treat the input token sequences during the encoding phase as if the requests had the same lengths as the input token sequence with the shortest length (i.e., least number of tokens). However, this requires the inference system 130 to discard output tokens generated for requests with longer input lengths at subsequent iterations since the original input tokens received for the request have to be substituted in the respective elements. This results in inefficiency and waste

of unnecessary resources because an inference system processes the input token sequences for longer requests separately in multiple iterations instead of a single iteration. Another workaround may be to treat the input token sequences as if the requests had the same lengths as the input token sequence with the longest length (i.e., largest number of tokens). However, this requires the inference system to increase the size of the input tensors to match the request with the longest token sequence (and pad the remaining elements). This results in redundant computation since the inference system processes a larger input tensor throughout the process of executing the transformer model. For example, in some cases, the number of iterations required to process a batch of requests using the method of FIGs. 2A-2B may be significantly larger than the number of iterations to process each request individually. [0050] Thus, in one embodiment, the inference system 130 trains and applies a machine-learning transformer model to a batch of requests with variable input length or target length or internal state length using a selective batching method. In the selective batching method, the inference system 130 selectively batches a subset of operations in the transformer model but separately processes requests individually for a subset of operations in the transformer model that are not compatible with batching. In one embodiment, the operation to be processed separately is the attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system 130 can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing problems that occur in workarounds.

2. More Like Text: ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. More Like Text: machine learning

Filters: Same as Query 12

14 2022-01-31 22:04 UTC | ~2253 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: Same as Query 7

Concept Modifiers (5):

5. More Like Text: methods of batching for transformer models often require that the length of data for multiple requests in a batch be the same to be processed. Thus, it may not be feasible to process a batch of requests with variable lengths or workarounds addressing this problem may result in using more resources compared to processing each request individually.

4. More Like Text: decoder

3. More Like Text: In many instances, the inputs for requests or the internal states for the requests are variable in length and may be difficult to batch using, for example, the batching method in FIGs. 2A-2B. For example, the number of input tokens between different NLP requests are highly likely to vary in length depending on the input text and the type of request. For example, the length of a first query request may be different from the length of a second query request since different users synthesize language differently. However, methods of batching transformer models, such as the method described in conjunction with FIGs. 2A-2B, require that the lengths of input tensors for the batch be the same since the input tensors have to be coalesced into a single concatenated tensor for an operation. Thus, if the lengths of input tensors are different for multiple requests, it may not be feasible to process the requests as a batch using the batching method of FIGs. 2A-2B. [0049] In addition, a workaround addressing this problem is to treat the

input token sequences during the encoding phase as if the requests had the same lengths as the input token sequence with the shortest length (i.e., least number of tokens). However, this requires the inference system 130 to discard output tokens generated for requests with longer input lengths at subsequent iterations since the original input tokens received for the request have to be substituted in the respective elements. This results in inefficiency and waste of unnecessary resources because an inference system processes the input token sequences for longer requests separately in multiple iterations instead of a single iteration. Another workaround may be to treat the input token sequences as if the requests had the same lengths as the input token sequence with the longest length (i.e., largest number of tokens). However, this requires the inference system to increase the size of the input tensors to match the request with the longest token sequence (and pad the remaining elements). This results in redundant computation since the inference system processes a larger input tensor throughout the process of executing the transformer model. For example, in some cases, the number of iterations required to process a batch of requests using the method of FIGs. 2A-2B may be significantly larger than the number of iterations to process each request individually. [0050] Thus, in one embodiment, the inference system 130 trains and applies a machine-learning transformer model to a batch of requests with variable input length or target length or internal state length using a selective batching method. In the selective batching method, the inference system 130 selectively batches a subset of operations in the transformer model but separately processes requests individually for a subset of operations in the transformer model that are not compatible with batching. In one embodiment, the operation to be processed separately is the attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system 130 can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing problems that occur in workarounds.

2. More Like Text: ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. More Like Text: machine learning

Filters: Same as Query 12

15 2022-01-31 22:10 UTC | ~6225 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: Same as Query 7

Concept Modifiers (6):

6. More Like Text: batching requests with variable lengths

5. More Like Text: methods of batching for transformer models often require that the length of data for multiple requests in a batch be the same to be processed. Thus, it may not be feasible to process a batch of requests with variable lengths or workarounds addressing this problem may result in using more resources compared to processing each request individually.

4. More Like Text: decoder

3. More Like Text: In many instances, the inputs for requests or the internal states for the requests are variable in length and may be difficult to batch using, for example, the batching method in FIGs. 2A-2B. For example, the number of input tokens between different NLP requests are highly likely to vary in length depending on the input text and the type of request. For example, the length of a first query request may be different from the length of a second query

request since different users synthesize language differently. However, methods of batching transformer models, such as the method described in conjunction with FIGs. 2A-2B, require that the lengths of input tensors for the batch be the same since the input tensors have to be coalesced into a single concatenated tensor for an operation. Thus, if the lengths of input tensors are different for multiple requests, it may not be feasible to process the requests as a batch using the batching method of FIGs. 2A-2B. [0049] In addition, a workaround addressing this problem is to treat the input token sequences during the encoding phase as if the requests had the same lengths as the input token sequence with the shortest length (i.e., least number of tokens). However, this requires the inference system 130 to discard output tokens generated for requests with longer input lengths at subsequent iterations since the original input tokens received for the request have to be substituted in the respective elements. This results in inefficiency and waste of unnecessary resources because an inference system processes the input token sequences for longer requests separately in multiple iterations instead of a single iteration. Another workaround may be to treat the input token sequences as if the requests had the same lengths as the input token sequence with the longest length (i.e., largest number of tokens). However, this requires the inference system to increase the size of the input tensors to match the request with the longest token sequence (and pad the remaining elements). This results in redundant computation since the inference system processes a larger input tensor throughout the process of executing the transformer model. For example, in some cases, the number of iterations required to process a batch of requests using the method of FIGs. 2A-2B may be significantly larger than the number of iterations to process each request individually. [0050] Thus, in one embodiment, the inference system 130 trains and applies a machine-learning transformer model to a batch of requests with variable input length or target length or internal state length using a selective batching method. In the selective batching method, the inference system 130 selectively batches a subset of operations in the transformer model but separately processes requests individually for a subset of operations in the transformer model that are not compatible with batching. In one embodiment, the operation to be processed separately is the attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system 130 can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing problems that occur in workarounds.

2. More Like Text: ABSTRACT OF THE DISCLOSURE An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length (BENEFIT).

1. More Like Text: machine learning

Filters: Same as Query 12

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	1	of	2	Attorney Docket Number	39192-50678/US

U.S. PATENT DOCUMENTS				
Examiner Initials*	Cite No. ¹	Document No. Number – Kind Code ² (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document
	A1	US-2021/0034335 A1	02-04-2021	Svyalkovskly et al.
	A2	US-2021/0357210 A1	11-18-2021	Clement et al.
	A3	US-2022/0066747 A1	03-03-2022	Drain et al.
	A4	US-2022/0067513 A1	03-03-2022	Stevens et al.

FOREIGN PATENT DOCUMENTS					
Examiner Initials*	Cite No. ¹	Foreign Patent Document Country Code ³ – Number ⁴ Kind Code ⁵ (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	T ⁶
	B1	N/A			

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS				
Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published		T ⁶
	C1	DOSHI, Ketan, "Transformers Explained Visually (Part 1): Overview of Functionality", Dec.13, 2020, <towardsdatascience.com> (Year: 2020), 16 pages.		
	C2	DOSHI, Ketan, "Transformers Explained Visually (Part 2): How it works, step-by-step", Jan.2, 2021, <towardsdatascience.com> (Year: 2021), 23 pages.		
	C3	DOSHI, Ketan, "Transformers Explained Visually (Part 3): Multi-head Attention deep dive", Jan.16, 2021, <towardsdatascience.com> (Year: 2021), 20 pages.		
	C4	DOSHI, Ketan, "Transformers Explained Visually (Part 4): Not Just How, but Why They Work So Well", June 2, 2021, <towardsdatascience.com> (Year: 2021), 17 pages.		
	C5	VASWANI et al., "Attention Is All You Need", Dec. 6, 2017, arXiv, <https://arxiv.org/abs/1706.03762> (Year: 2017), pages 1-15.		

EXAMINER SIGNATURE	
Examiner Signature	
Date Considered	
*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.	
¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.	
⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	2	of	2	Attorney Docket Number	39192-50678/US

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☒ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

☐ After payment of the issue fee. Applicant requests that the information contained in this Information Disclosure Statement be placed in the file according to 37 CFR § 1.97(i), although the information may not be considered by the USPTO.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Jae Yeon Baek/	Date (YYYY-MM-DD)	2022-04-12
Name/Print	Jae Yeon Baek	Registration Number	78,258

Electronic Patent Application Fee Transmittal

Application Number:	17542193			
Filing Date:	03-Dec-2021			
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks			
First Named Inventor/Applicant Name:	Gyeongin Yu			
Filer:	Jae Yeon Baek/Julie Eng			
Attorney Docket Number:	39192-50678/US			
Filed as Small Entity				
Filing Fees for Utility under 35 USC 111(a)				
Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Miscellaneous:				
SUBMISSION- INFORMATION DISCLOSURE STMT	2806	1	130	130
Total in USD (\$)				130

Electronic Acknowledgement Receipt

EFS ID:	45459972
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Yeon Baek/Julie Eng
Filer Authorized By:	Jae Yeon Baek
Attorney Docket Number:	39192-50678/US
Receipt Date:	12-APR-2022
Filing Date:	03-DEC-2021
Time Stamp:	23:13:47
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	\$ 130
RAM confirmation Number	E20224BN14104250
Deposit Account	
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

File Listing:					
Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Information Disclosure Statement (IDS) Form (SB08)	50678_US_Information_Disclosure_Statement.pdf	222487	no	2
			bcfd2d290e89d0e420c10a875b7704da72acdc6a		
Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					
2	Non Patent Literature	DOSHI_Transformers_Explained_Visually_Part_1_Overview_of_Functionality.pdf	1789472	no	16
			ef1f7b8b5b18bb90a97de954270d97d48af1a8c1		
Warnings:					
Information:					
3	Non Patent Literature	DOSHI_Transformers_Explained_Visually_Part_2_How_it_works_step-by-step.pdf	2436140	no	23
			8e7d912fe3aac7ba1331ea6e5443315b8633118a		
Warnings:					
Information:					
4	Non Patent Literature	DOSHI_Transformers_Explained_Visually_Part_3_Multi-head_Attention_deep_dive.pdf	2009778	no	20
			f5877ca290d5af9c23ddf69aaa77ad6e96f7e19d		
Warnings:					
Information:					
5	Non Patent Literature	DOSHI_Transformers_Explained_Visually_Part_4_Not_Just_How_but_Why_They_Work_So_Well.pdf	1776020	no	17
			bf0b7631bfec5f5c18c5dd79dce14800d0d62146		
Warnings:					
Information:					
6	Non Patent Literature	VASWANI_Attention_Is_All_You_Need.pdf	837681	no	15
			280288392f8a7713c716050da3e74af68bab4759		
Warnings:					

Information:					
7	Fee Worksheet (SB06)	fee-info.pdf	37955	no	2
			db89709480b5f003d25ab9706ce53fb35b426348		
Warnings:					
Information:					
Total Files Size (in bytes):				9109533	
This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503. <u>New Applications Under 35 U.S.C. 111</u> If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application. <u>National Stage of an International Application under 35 U.S.C. 371</u> If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course. <u>New International Application Filed with the USPTO as a Receiving Office</u> If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.					



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026
758	7590	05/02/2022	EXAMINER	
FENWICK & WEST LLP			TANG, KENNETH	
SILICON VALLEY CENTER			ART UNIT	
801 CALIFORNIA STREET			PAPER NUMBER	
MOUNTAIN VIEW, CA 94041			2199	
			NOTIFICATION DATE	DELIVERY MODE
			05/02/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

PTOC@Fenwick.com

<i>Applicant-Initiated Interview Summary</i>	Application No. 17/542,193	Applicant(s) Yu et al.		
	Examiner KENNETH TANG	Art Unit 2199	AIA (First Inventor to File) Status Yes	Page 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	Title	Type
KENNETH TANG	Primary Examiner	Telephonic
Claire Baek	Attorney of Record	
Jae Won Song	Attorney of Record	

Date of Interview: 27 April 2022

Issues Discussed:

35 U.S.C. 103

It was agreed that the proposed claim amendment in addition to the idea of, the second batch of requests has a different size than the previous one, would overcome the current prior art rejections. However, allowability could not be determined at this time.

Proposed Amendment(s)

It was agreed that the proposed claim amendment in addition to the idea of, the second batch of requests has a different size than the previous one, would overcome the current prior art rejections. However, allowability could not be determined at this time.

/KENNETH TANG/ Primary Examiner, Art Unit 2199	
Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04 Please further see: MPEP 713.04 Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b) 37 CFR § 1.2 Business to be transacted in writing	

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/542,193	Filing Date 12/03/2021	☐ To be Mailed
---	--	---------------------------	---------------------------------------

ENTITY: ☐ LARGE ☒ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA	RATE (\$)	FEE (\$)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*	x \$50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*	x \$240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is \$310 (\$155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).			
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)	RATE (\$)	ADDITIONAL FEE (\$)
AMENDMENT	05/03/2022	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		
	Total (37 CFR 1.16(i))	* 18	Minus	** 20	= 0	x \$50 =	0
	Independent (37 CFR 1.16(h))	* 2	Minus	*** 3	= 0	x \$240 =	0
	☐ Application Size Fee (37 CFR 1.16(s))						
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))						
						TOTAL ADD'L FEE	0
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		
	Total (37 CFR 1.16(i))	*	Minus	**	=	x \$0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x \$0 =	
	☐ Application Size Fee (37 CFR 1.16(s))						
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))						
						TOTAL ADD'L FEE	
* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.						SLIE	
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".						/ADRIENE D. SELLMAN/	
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".							
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.							

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

FIRST NAMED INVENTOR: Gyeongin Yu

APPLICATION NO.: 17/542,193

TITLE: Dynamic Batching for Inference System for Transformer-Based Generation Tasks

FILING DATE: December 3, 2021

EXAMINER: Kenneth Tang

GROUP ART UNIT: 2199

CONFIRMATION NO.: 1026

ATTY. DKT. NO.: 39192-50678/US

CERTIFICATE OF TRANSMISSION			
I hereby certify under 37 C.F.R. § 1.8(a)(ii) that this correspondence is being deposited or transmitted in accordance with 37 C.F.R. § 1.8(a)(i) on or before the local date shown below.			
Signature:	/Jae Yeon Baek/		
Typed or Printed Name:	Jae Yeon Baek, Reg. No. 78,258	Dated:	May 3, 2022

RESPONSE AND AMENDMENT A

The following is in response to the Office Action mailed February 9, 2022.

AMENDMENTS TO THE SPECIFICATION

Please amend the Abstract of the Disclosure as shown below:

An inference system applies a machine-learning transformer model to a batch of requests with variable input length or variable target length or variable internal ~~[[sate]]~~ state length by selectively batching a subset of operations in the transformer model but processing requests in the batch individually for a subset of operations in the transformer model. In one embodiment, the operation to be processed individually is an attention operation of an encoder or a decoder of the transformer model. By selective batching, the inference system can allow batching operations to be performed for a batch of requests with variable input or target length or internal state length to utilize the parallel computation capabilities of hardware accelerators while preventing unnecessary computations that occur for workarounds that restrain the data of a batch of requests to a same length.

AMENDMENTS TO THE CLAIMS

All pending claims, along with a current status of each, are reproduced below. If indicated, please amend the claims accordingly.

1. (Currently Amended) A method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, comprising:
receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders;
scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine;
generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests;
receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens;
scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests, wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request; and

generating, by the execution engine, a second set of output tokens by applying the transformer model to ~~[[a]]~~ the second set of inputs for the second batch ~~of requests including the sequence of input tokens for the new request.~~

2. (Currently Amended) The method of claim 1, further comprising:
responsive to determining that a request in the first batch of requests has been completed,
providing output tokens generated for the completed request to a client device as a response to the request, and
wherein the second batch of requests includes at least one of the remaining requests from the one or more requests and the new request.

3. (Currently Amended) The method of claim 2, wherein the request is associated with a cache memory ~~[[cache]]~~ in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated cache memory ~~[[cache]]~~ for the request in the execution engine.

4. (Currently Amended) The method of claim 1, wherein ~~the second set of inputs include the sequence of input tokens for the new request and~~ the input for the at least one request ~~is~~ an output token from the first set of output tokens for the at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request.

5. (Original) The method of claim 4,
wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and

wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request.

6. (Original) The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request.

7. (Original) The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests.

8. (Currently Amended) The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a ~~[[tenor]]~~ tensor processing unit (TPU).

9. (Original) The method of claim 1, wherein each token in the sequence of input tokens represents a text unit.

10. (Currently Amended) A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising:

receiving, by a serving system, one or more requests for execution, the serving system

including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders;

scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine;

generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests;

receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens;

scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests, wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request; and

generating, by the execution engine, a second set of output tokens by applying the transformer model to ~~[[a]] the~~ second set of inputs for the second batch of ~~requests including the sequence of input tokens for the new request.~~

11. (Currently Amended) The non-transitory computer-readable storage medium of claim 10, the operations further comprising:

responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes at least one of the remaining requests from the one or more requests and the new request.

12. (Currently Amended) The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a cache memory [[cache]] in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated cache memory [[cache]] for the request in the execution engine.

13. (Currently Amended) The non-transitory computer-readable storage medium of claim 10, wherein ~~the second set of inputs include the sequence of input tokens for the new request and~~ the input for the at least one request is at least one output token from the first set of output tokens for the at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request.

14. (Original) The non-transitory computer-readable storage medium of claim 13, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and

wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request.

15. (Original) The non-transitory computer-readable storage medium of claim 14, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request.

16. (Original) The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests.

17. (Currently Amended) The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a ~~[[tenor]]~~ tensor processing unit (TPU).

18. (Original) The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

REMARKS

Claims 1-18 were pending in this application, and claims 1 and 10 were the independent claims. In the Office Action dated February 9, 2022, claims 1-4, 7-13, and 16-18 were rejected and claims 5-6 and 14-15 were objected to.

With this amendment, claims 1-4, 8, 10-13, and 17 are amended; no claims are canceled; and no claims are added. Support for the amendments appears at, for example, paragraph [0099] of the specification-as-filed.

In view of the amendments and remarks that follow, reconsideration of the pending claims is respectfully requested.

STATEMENT OF SUBSTANCE OF INTERVIEW

Applicant's representative Jae Yeon (Claire) Baek (Reg. No. 78,258) and Jae Won Song (Reg. No. 59,070) conducted a telephonic interview with Examiner Kenneth Tang on April 27, 2022. During the interview, the pending claims and the rejections to those claims under 35 U.S.C. § 102/103 in light of references Chung and Aarts were discussed. An agreement was reached that amendments similar to those proposed herein would overcome the current rejections. Examiner also indicated further search may be required upon submission of the response.

OBJECTIONS TO THE SPECIFICATION

The abstract of the disclosure was objected to. The abstract is hereby amended to correct the typographical error.

OBJECTIONS TO THE CLAIMS

Claims 8 and 17 are objected to because of informalities. Claims 8 and 17 are hereby amended to correct the typographical errors.

REJECTIONS UNDER 35 U.S.C. § 102 and/or 35 U.S.C. § 103

Claims 1–4, 7–13, and 16–18 were rejected under 35 U.S.C. § 103 as unpatentable over U.S. Patent App. Pub. No. 2021/0192314 (“Aarts”) in view of U.S. Patent No. 10,846,096 (“Chung”).

Claim 1, as amended, recites the feature of “scheduling, by the scheduler, **a second batch of requests additionally including the new request for execution on the execution engine**, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests, **wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request** . . . and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request.”

Specifically, the method of claim 1 recites a method of dynamically executing batches of requests for a machine-learning transformer model:

- a serving system receives one or more requests for execution,
- a scheduler schedules a batch of requests including the one or more requests,
- an execution engine generates a first set of output tokens by applying the transformer model to a set of inputs for the batch of requests,
- a request processor receives a new request from a client device,

- the scheduler schedules a second batch of requests additionally including the new request for execution on the execution engine, wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request, and
- the execution engine generates a second set of output tokens by applying the transformer model to the second set of inputs for the second batch of requests.

Thus, as discussed during the interview, in the method of claim 1, the second batch of requests is modified to include a new request in addition to the one or more requests of the [first] batch of requests.

In existing batching methods for transformer models, it is difficult to modify a batch of requests once it has started to process on an execution engine, since the length of the inputs or internal states are the same across the requests in the batch. *See* specification, ¶ [0009]. However, by performing the dynamic batching method described in conjunction with FIGs. 5A-5D of the application with a transformer architecture with selective batching (e.g., described in conjunction with FIGs. 3A-3B), an inference system can modify a batch being processed and update the batch between iterations as new requests are received, even if a length of the input for the new request is different from a length of the input for another request in the batch.

Chung generally discloses a method for “the batching of requests for a specific service or to be processed using a specific model.” Chung, col. 3, ll. 22-25. For a given batch request, the memory needed for processing in the GPU can be pre-allocated before the service starts. Chung, col. 3, ll. 49-51. However, Chung does not describe batching in the context of machine-learned transformer models. In addition, Chung does not disclose applying a transformer model to a

[first] batch of requests on a GPU and subsequently scheduling a second batch of requests that is modified to include a new request in addition to the first batch of requests on the GPU, in which a length of the sequence of input tokens for a new request is different from a length of an input for at least one request other than the new request in the batch.

Another portion of Chung discloses that “[o]nce the requests are pulled from the queue, the queue can be flushed if necessary such that the next received request for the relevant service or model will be the first new request in the queue for the next batch.” Chung, col. 3, ll. 40-46. However, Chung describes scheduling the next received request **in a queue for a next batch**, but does not disclose modifying the existing batch on a GPU by additionally including the next received request. In other words, Chung appears to describe storing the next received request in the queue after existing requests are pulled from the queue, such that the next received request becomes the first new request for **a new batch of requests**.

Therefore, Chung fails to disclose or suggest the feature of “scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests, wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request . . . and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request,” as recited in claim 1.

Aarts generally discloses making one or more API calls for executing a recurrent neural network. FIG. 6 of Aarts discloses a ragged recurrent neural network in which each batch 606 of

an input tensor 602 has a different-length sequence 608. However, Aarts also does not disclose batching in the context of machine-learned transformer models. In addition, Aarts also does not disclose applying a transformer model to a first batch and subsequently scheduling a second batch of requests that is modified to include a new request in addition to the first batch of requests, in which a length of the sequence of input tokens for a new request is different from a length of an input for at least one request other than the new request in the batch.

Therefore, Aarts also fails to disclose or suggest the feature of “scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests, wherein in a second set of inputs for the second batch of requests, a length of the sequence of input tokens for the new request is different from a length of an input for at least one request other than the new request . . . and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request,” as recited in claim 1.

The other cited references do not cure the deficiencies of Chung and Aarts.

Thus, claim 1 and its dependent claims are patentably distinguishable over the cited references. Claim 10 recites similar features as claim 1. Thus, claim 10 and its dependent claims are also patentably distinguishable over the cited references.

Applicant respectfully requests the rejections be withdrawn.

CONCLUSION

Based on the foregoing, the application is in condition for allowance of all claims, and a Notice of Allowance is respectfully requested. If the examiner believes for any reason direct contact would help advance the prosecution of this case to allowance, the examiner is encouraged to telephone the undersigned at the number given below.

If extensions of time are necessary to prevent abandonment of this application, then such extension of time are hereby petitioned under 37 CFR 1.136(a), and any fees required therefore are hereby authorized to be charged to our Deposit Account 19-2555.

Respectfully submitted,

Dated: May 3, 2022

/Jae Yeon Baek/
Jae Yeon Baek, Reg. No. 78,258
FENWICK & WEST LLP
801 California Street
Mountain View, CA 94041
Phone: (650) 335-7142

Electronic Acknowledgement Receipt

EFS ID:	45620817
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Yeon Baek
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	03-MAY-2022
Filing Date:	03-DEC-2021
Time Stamp:	14:23:04
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no
------------------------	----

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		50678_US_Amendment_A.pdf	166397	yes	14
			86150c72c0edfc6589a92e23eb79406ba7a2d023		

	Multipart Description/PDF files in .zip description		
	Document Description	Start	End
	Amendment/Request for Reconsideration-After Non-Final Rejection	1	1
	Specification	2	2
	Claims	3	8
	Applicant Arguments/Remarks Made in an Amendment	9	14

Warnings:

Information:

Total Files Size (in bytes):	166397
-------------------------------------	--------

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

758 7590 05/11/2022
FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

EXAMINER	
TANG, KENNETH	
ART UNIT	PAPER NUMBER
2199	

DATE MAILED: 05/11/2022

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026

TITLE OF INVENTION: Dynamic Batching for Inference System for Transformer-Based Generation Tasks

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	\$600	\$0.00	\$0.00	\$600	08/11/2022

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the ENTITY STATUS shown above. If the ENTITY STATUS is shown as SMALL or MICRO, verify whether entitlement to that entity status still applies.

If the ENTITY STATUS is the same as shown above, pay the TOTAL FEE(S) DUE shown above.

If the ENTITY STATUS is changed from that shown above, on PART B - FEE(S) TRANSMITTAL, complete section number 5 titled "Change in Entity Status (from status indicated above)".

For purposes of this notice, small entity fees are 1/2 the amount of undiscounted fees, and micro entity fees are 1/2 the amount of small entity fees.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Maintenance fees are due in utility patents issuing on applications filed on or after Dec. 12, 1980. It is patentee's responsibility to ensure timely payment of maintenance fees when due. More information is available at www.uspto.gov/PatentMaintenanceFees.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via EFS-Web.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

758 7590 05/11/2022
FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via EFS-Web or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026

TITLE OF INVENTION: Dynamic Batching for Inference System for Transformer-Based Generation Tasks

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	\$600	\$0.00	\$0.00	\$600	08/11/2022

EXAMINER	ART UNIT	CLASS-SUBCLASS
TANG, KENNETH	2199	718-101000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

☐ Change of correspondence address (or Change of Correspondence Address form PTO/AIA/122 or PTO/SB/122) attached.

☐ "Fee Address" indication (or "Fee Address" Indication form PTO/AIA/47 or PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list

(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,

1 _____

(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

2 _____

3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE

(B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ☐ Individual ☐ Corporation or other private group entity ☐ Government

4a. Fees submitted: ☐ Issue Fee ☐ Publication Fee (if required) ☐ Advance Order - # of Copies _____

4b. Method of Payment: (Please first reapply any previously paid fee shown above)

☐ Electronic Payment via EFS-Web ☐ Enclosed check ☐ Non-electronic payment by credit card (Attach form PTO-2038)

☐ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. _____

5. Change in Entity Status (from status indicated above)

☐ Applicant certifying micro entity status. See 37 CFR 1.29

☐ Applicant asserting small entity status. See 37 CFR 1.27

☐ Applicant changing to regular undiscounted fee status.

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.

NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.

NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature _____

Date _____

Typed or printed name _____

Registration No. _____



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026
758	7590	05/11/2022	EXAMINER	
FENWICK & WEST LLP SILICON VALLEY CENTER 801 CALIFORNIA STREET MOUNTAIN VIEW, CA 94041			TANG, KENNETH	
			ART UNIT	PAPER NUMBER
			2199	
DATE MAILED: 05/11/2022				

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b) (Applications filed on or after May 29, 2000)

The Office has discontinued providing a Patent Term Adjustment (PTA) calculation with the Notice of Allowance.

Section 1(h)(2) of the AIA Technical Corrections Act amended 35 U.S.C. 154(b)(3)(B)(i) to eliminate the requirement that the Office provide a patent term adjustment determination with the notice of allowance. See Revisions to Patent Term Adjustment, 78 Fed. Reg. 19416, 19417 (Apr. 1, 2013). Therefore, the Office is no longer providing an initial patent term adjustment determination with the notice of allowance. The Office will continue to provide a patent term adjustment determination with the Issue Notification Letter that is mailed to applicant approximately three weeks prior to the issue date of the patent, and will include the patent term adjustment on the patent. Any request for reconsideration of the patent term adjustment determination (or reinstatement of patent term adjustment) should follow the process outlined in 37 CFR 1.705.

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

OMB Clearance and PRA Burden Statement for PTOL-85 Part B

The Paperwork Reduction Act (PRA) of 1995 requires Federal agencies to obtain Office of Management and Budget approval before requesting most types of information from the public. When OMB approves an agency request to collect information from the public, OMB (i) provides a valid OMB Control Number and expiration date for the agency to display on the instrument that will be used to collect the information and (ii) requires the agency to inform the public about the OMB Control Number's legal significance in accordance with 5 CFR 1320.5(b).

The information collected by PTOL-85 Part B is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450. Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Notice of Allowability	Application No. 17/542,193	Applicant(s) Yu et al.	
	Examiner KENNETH TANG	Art Unit 2199	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to the Amendment/Remarks on 5/3/22.
☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on ____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on ____; the restriction requirement and election have been incorporated into this action.

3. ☒ The allowed claim(s) is/are 1-18. As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to PPHfeedback@uspto.gov.

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a) ☐ All b) ☐ Some* c) ☐ None of the:

1. ☐ Certified copies of the priority documents have been received.

2. ☐ Certified copies of the priority documents have been received in Application No. ____.

3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

* Certified copies not received: ____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date ____.

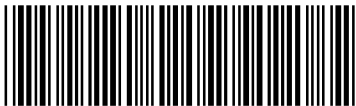
Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892) 2. ☒ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date <u>4/12/22</u> . 3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____. 4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.	5. ☐ Examiner's Amendment/Comment 6. ☐ Examiner's Statement of Reasons for Allowance 7. ☐ Other _____.
---	---

/KENNETH TANG/ Primary Examiner, Art Unit 2199	
---	--

Issue Classification 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

CPC						
Symbol					Type	Version
G06F	/	9	/	4881	F	2013-01-01
G06N	/	5	/	04	I	2013-01-01
G06N	/	20	/	00	I	2019-01-01
G06F	/	9	/	5016	I	2013-01-01
G06N	/	3	/	0454	A	2013-01-01
G06N	/	3	/	08	A	2013-01-01

CPC Combination Sets					
Symbol				Type	Version
	/		/		

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	18	
/KENNETH TANG/ Primary Examiner, Art Unit 2199	05 May 2022	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	3A

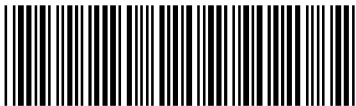
<i>Issue Classification</i> 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

INTERNATIONAL CLASSIFICATION			
CLAIMED			
G06F	/	9	/ 46
NON-CLAIMED			
/		/	

US ORIGINAL CLASSIFICATION	
CLASS	SUBCLASS
718	101

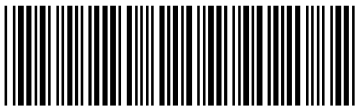
CROSS REFERENCES(S)						
CLASS	SUBCLASS (ONE SUBCLASS PER BLOCK)					

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	18	
/KENNETH TANG/ Primary Examiner, Art Unit 2199	05 May 2022	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	3A

<i>Issue Classification</i> 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

☒ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47															
CLAIMS															
Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original
	1		10												
	2		11												
	3		12												
	4		13												
	5		14												
	6		15												
	7		16												
	8		17												
	9		18												

NONE (Assistant Examiner) _____ (Date) _____		Total Claims Allowed: 18	
/KENNETH TANG/ Primary Examiner, Art Unit 2199 (Primary Examiner) _____ (Date) 05 May 2022		O.G. Print Claim(s) 1	O.G. Print Figure 3A

<i>Search Notes</i> 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

CPC - Searched*		
Symbol	Date	Examiner
G06F9/4881, G06N20/00, G06F9/5016, G06N5/04, G06N3/0454, G06N3/08	02/01/2022	KT

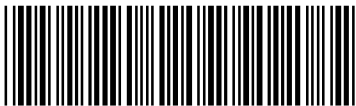
CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner
718	1	02/01/2022	KT

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
PE2E search	02/01/2022	KT
Inventor search	02/01/2022	KT
IP.com search	02/01/2022	KT
Google Patents and NPL Scholar search	02/01/2022	KT
PE2E updated search	05/05/2022	KT
IP.com NPL search	05/05/2022	KT

/KENNETH TANG/ Primary Examiner, Art Unit 2199	
---	--

<i>Search Notes</i> 	Application/Control No. 17/542,193	Applicant(s)/Patent Under Reexamination Yu et al.
	Examiner KENNETH TANG	Art Unit 2199

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner
G06F	9/4881, 9/5016	05/05/2022	KT
G06N	20/00, 5/04, 3/0454, 3/08	05/05/2022	KT
	US Pat and PG Pub search	05/05/2022	KT

/KENNETH TANG/ Primary Examiner, Art Unit 2199	
---	--

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L105	0	17/542193.app.	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L106	1	"20210192314".pn.	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L107	1	("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/05/05 04:02 PM
L108	1	"20210192314".pn. AND (token text)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L109	1	"20210192314".pn. AND (token text sequence)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L110	478010	(718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L111	6639	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L112	23	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4))	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL,	OR	ON	ON	2022/05/05 04:02 PM

L113	23	AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. ((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L114	5	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L115	5	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L116	4	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

L117	1	AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/05/05 04:02 PM
L118	4	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key AND sequence	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L119	2	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND (batch\$4 AND (ai "artificial intelligence" (machine ADJ learn\$4)) AND schedul\$4 AND (token decod\$4 transform\$4)).ab,ti,clm. AND @ad<"20211203" AND (text word) AND key AND sequence	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L120	5435	AND tensor	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L121	2354	AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

L122	813	AND schedul\$4 ((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L123	753	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L124	266	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L125	47	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu) AND length	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L126	45	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND batch\$4 AND decod\$4 AND @ad<"20211203" AND schedul\$4 AND tensor AND (gpu tpu) AND (length near5 (tensor key cache request different))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

		(tensor key cache request different)) AND decoder					
L127	1	"20210192314".pn. AND (tpu gpu)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L128	1	"20210192314".pn. AND length	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L129	1	"20210192314".pn. AND (key AND cache)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L130	1	"20210192314".pn. AND (tensor WITH cache)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L131	1	"20210192314".pn. AND (dynamic\$4 (batch\$4 WITH request) (learn\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L132	1	"20210192314".pn. AND batch\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L133	1	"20210192314".pn. AND decod\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L134	7	("20200285977" "20200285912" "20190220703" "20210240825" "20190138934" "20210240826" "20210185066").pn. AND ((variable different) near5 length)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L135	1	"20210192314".pn. AND cache	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L136	1	"20210192314".pn. AND (cache SAME (key tensor length value request batch))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L137	1	"20210192314".pn. AND (gpu tpc tcu)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L138	1	"20210192314".pn. AND (token sequence text)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L139	1	"20210192314".pn. AND insufficien\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L140	1	"20210192314".pn. AND availab\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L141	1	("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/05/05 04:02 PM
L142	1	"20210192314".pn. AND transform\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L143	1	"20210192314".pn. AND availab\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L144	1	"20210192314".pn.	(US-PGPUB)	OR	ON	ON	2022/05/05

L145	0	AND weight "20210192314".pn. AND (decoder SAME (model learn\$4 ai))	(US-PGPUB)	OR	ON	ON	04:02 PM 2022/05/05 04:02 PM
L146	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L147	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L148	0	"20210192314".pn. AND (decoder SAME (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L149	1	"20210192314".pn. AND (decoder AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L150	1	"20210192314".pn. AND (schedul\$4 AND (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L151	1	"20210192314".pn. AND (schedul\$4 SAME (model learn\$4 ai translat\$4 transform\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L152	1	"20210192314".pn. AND (schedul\$4 WITH (batch request))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L153	1	"20210192314".pn. AND batch	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L154	1	"20210192314".pn. AND (output WITH (token model transform\$4 translat\$4 learn\$4 batch request))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L155	1	"20210192314".pn. AND length	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L156	1	"20210192314".pn. AND ((new second) WITH (request batch))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L157	1	"20210192314".pn. AND (split\$4 concatenat\$4)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L158	1	"20210192314".pn. AND cache	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L159	0	"20190251418".pn. AND ((memory resource) WITH (sufficent insufficient enough availab\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L160	0	"20190251418".pn.	(US-PGPUB)	OR	ON	ON	2022/05/05

		AND ((memory resource) SAME (sufficient insufficient enough availab\$4))					04:02 PM
L161	1	"20190251418".pn. AND ((sufficient insufficient enough availab\$4))	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L162	1	"20210192314".pn. AND remain\$4	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L163	1	"20210192314".pn. AND (second new)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L164	152	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L165	13	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L166	13	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L167	12	(G06F9/5016).cpc. AND (G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND decoder AND @ad<"20211203" AND ((availab\$4 sufficien\$4 insufficien\$4 threshold) near4 (capacity memory resource))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L168	14	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn.	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

L169	14	AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight sufficien\$4 insufficien\$4 threshold availab\$4 length split\$4 concaten\$4 cache) ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L170	0	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L171	6	AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND G06F9/5016.pn. ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

L172	6	"20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. ("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868" "20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. AND decoder	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L173	3	("20210192287" "20200364088" "20180314249" "20210303481" "20210326176" "20200293456" "20210390058" "20210142438" "20210191868"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM

		"20210349831" "20210279837" "11216717" "20200043123" "20210390460").pn. AND (decod\$4 normaliz\$4 learn\$4 batch\$4 weight length split\$4 concaten\$4) AND ((memory resource cache) near4 (availab\$4 sufficien\$4 insufficien\$4 threshold)) AND (batch\$4 schedul\$4 tensor token tranform\$4 model learn\$4).ab,ti,clm. AND decoder AND keys					
L174	4	("10846096" "20210334134" "20210232399" "9652287").pn. AND (new WITH (request batch))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L175	1	("10846096").pn. AND (new WITH (request batch)) AND (memory resource capacity) WITH (availab\$4 sufficien\$4 insufficien\$4 threshold)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L176	1	("10846096").pn. AND (new WITH (request batch)) AND ((memory resource capacity) WITH (availab\$4 sufficien\$4 insufficien\$4 threshold))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L177	1	"20210192314".pn. AND new	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L178	1	"20210192314".pn. AND (availab\$4 sufficien\$4 insufficien\$4 threshold)	(US-PGPUB)	OR	ON	ON	2022/05/05 04:02 PM
L179	12381	((variable different) near4 length) AND L7	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR,	OR	ON	ON	2022/05/05 04:02 PM

L180	10067	((variable different) near4 length) AND L7 AND learn\$4	TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L181	925	((variable different) near4 length) AND L7 AND learn\$4 AND tensor	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L182	476	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L183	104	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL,	OR	ON	ON	2022/05/05 04:02 PM

L184	102	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203"	NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L185	1	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND (key near3 (cache tensor))	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L186	74	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND (decod\$4 transform\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L187	0	((variable different) near4 length) AND L7 AND learn\$4 AND tensor AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM

L188	471	(decod\$4 transform\$4) batch\$4 AND learn\$4 AND (decod\$4 transform\$4)	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L189	65	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L190	65	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203"	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L191	30	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu tcu)	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L192	38	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu normaliz\$4 tcu)	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L193	5	batch\$4 AND learn\$4 AND (decod\$4 transform\$4) AND L7 AND @ad<"20211203" AND (weight tensor gpu length tpu normaliz\$4 tcu) AND batch.ti.	(DERWENT)	OR	ON	ON	2022/05/05 04:02 PM
L194	30	(Artificial Intelligence) More Like: US-20210216870-A1 Text: (CA-2975251-C OR JP- 2020537204-A OR JP- 7009614-B2 OR WO- 2019238072-A1 OR US-10417562-B2 OR US-20200257979-A1 OR JP-2021192251-A OR US-10902319-B2 OR US-20200012942- A1 OR US- 20210357756-A1 OR JP-6935542-B2 OR US- 11281973-B2 OR US- 20200234127-A1 OR US-11308394-B2 OR AU-2019200309-B2 OR US-20200057924-A1 OR US-10628710-B2 OR KR-102204286-B1 OR US-20160217368-	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM

L195	1	A1 OR CN-107278310-A).did. ("2021/0192314").urpn. AND (PGPB USPT USOC).dbnm.	(US-PGPUB; USPAT; USOCR)	OR	ON	ON	2022/05/05 04:02 PM
L196	121923	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L197	6347	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm.	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L198	6343	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203"	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L199	760	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR,	OR	ON	ON	2022/05/05 04:02 PM

		tensor	HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)				
L200	18	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND tensor AND (variable near4 length)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L201	2	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND batch\$4.ab,ti,clm. AND @ad<"20211203" AND tensor AND (variable near4 length)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L202	0	(G06N3/0454 G06N5/04).cpc. AND (G06N20/00 G06N3/08).cpc. AND (batch WITH variable WITH length WITH request)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L203	3	(batch WITH variable WITH length WITH request)	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L204	25	(batch\$4 near3 request) AND (variable WITH length WITH (model learn\$4))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L205	25	batch\$4 WITH dynamic\$4 WITH (inference neural learn\$4) WITH (transform\$4 decod\$4)	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT,	OR	ON	ON	2022/05/05 04:02 PM

L206	24	L102 AND @ad<"20211203"	RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:02 PM
L207	1	L102 AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:02 PM
L208	9801	(transform\$4 near3 model) AND batch\$4	(US-PGPUB; USPAT; USOCR; FIT (AP, AT, AU, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DK, EA, EE, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, RU, SE, SG, SI, SK, SU, TH, TN, TR, TW, UA, VN, WO); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/05/05 04:06 PM
L209	446	(transform\$4 near3 model) AND batch\$4 AND ((length size) near5 (batch request))	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:06 PM
L210	364	(transform\$4 near3 model) AND batch\$4 AND ((length size) near5 (batch request)) AND learn\$4	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:06 PM
L211	361	(transform\$4 near3 model) AND batch\$4 AND ((length size) near5 (batch request)) AND learn\$4 AND @ad<"20211203"	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:07 PM

PE2E SEARCH - Search History (Interference)

Ref #	Hits	Search Query	DBs	Default	Plurals	British	Time Stamp
-------	------	--------------	-----	---------	---------	---------	------------

				Operator		Equivalents	
N1	113487	(718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:04 PM
N2	1	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND ((transform\$4 adj3 model) AND batch\$4 AND (length size)).clm.	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:05 PM
N3	5	((718/1.ccls.) (G06F9/4881 G06N20/00 G06F9/5016 G06N5/04 G06N3/0454 G06N3/08).cpc.) AND ((transform\$4 adj3 model) AND batch\$4).clm.	(US-PGPUB; USPAT)	OR	ON	ON	2022/05/05 04:05 PM



Search History

Report Run Date: 2022-05-05 UTC

Searches: 20

Markers: 0

Search #	Timestamp	Results	Page
1	2022-05-05 04:32 UTC	~58,580,867	3
2	2022-05-05 04:33 UTC	~17,620,602	4
3	2022-05-05 04:33 UTC	~15,895,460	5
4	2022-05-05 04:34 UTC	~132,506	5
5	2022-05-05 04:34 UTC	~37,444	5
6	2022-05-05 04:34 UTC	~12,474	7
7	2022-05-05 04:36 UTC	~1031	7
8	2022-05-05 04:36 UTC	~153	7
9	2022-05-05 04:41 UTC	4	7
10	2022-05-05 04:41 UTC	3	8
11	2022-05-05 04:45 UTC	0	8
12	2022-05-05 04:45 UTC	0	8
13	2022-05-05 04:45 UTC	0	9
14	2022-05-05 04:45 UTC	~559	9
15	2022-05-05 04:46 UTC	8	9
16	2022-05-05 04:48 UTC	6	9
17	2022-05-05 20:07 UTC	~16,007,094	10
18	2022-05-05 20:07 UTC	2,142,326	12
19	2022-05-05 20:08 UTC	2,046,771	14
20	2022-05-05 20:08 UTC	143,885	14



US Patents | US Designs | US Applications | EPO Patents | EPO Applications | China Patents | China Applications | Japan Patents | Japan Applications | Korea Patents | Korea Applications | WIPO Applications | Argentina Patents | Argentina Applications | Brazil Patents | Brazil Applications | Canada Patents | Canada Applications | Chile Patents | Chile Applications | Colombia Applications | Costa Rica Applications | Cuba Patents | Cuba Applications | Dominican Republic Applications | Ecuador Patents | Ecuador Applications | El Salvador Applications | Guatemala Applications | Honduras Applications | Mexico Patents | Mexico Applications | Nicaragua Patents | Panama Applications | Peru Applications | Trinidad & Tobago Patents | Uruguay Applications | Austria Patents | Austria Applications | Belarus Patents | Belgium Patents | Belgium Applications | Bosnia & Herzegovina Patents | Bosnia & Herzegovina Applications | Bulgaria Patents | Bulgaria Applications | Croatia Patents | Croatia Applications | Czech Republic Patents | Czech Republic Applications | Czechoslovakia Patents | Czechoslovakia Applications | Denmark Patents | Denmark Applications | Estonia Patents | Estonia Applications | EUIPO Patents | Finland Patents | Finland Applications | France Patents | France Applications | Germany Patents | Germany Applications | Great Britain Patents | Great Britain Applications | Greece Patents | Greece Applications | Hungary Patents | Hungary Applications | Iceland Patents | Iceland Applications | Ireland Patents | Ireland Applications | Italy Patents | Italy Applications | Latvia Patents | Latvia Applications | Lithuania Patents | Lithuania Applications | Luxembourg Patents | Malta Patents | Monaco Patents | Montenegro Patents | Montenegro Applications | Netherlands Patents | Netherlands Applications | Norway Patents | Norway Applications | Poland Patents | Poland Applications | Portugal Patents | Portugal Applications | Republic of Moldova Patents | Republic of Moldova Applications | Romania Patents | Romania Applications | San Marino Patents | San Marino Applications | Serbia Patents | Serbia Applications | Slovakia Patents | Slovakia Applications | Slovenia Patents | Spain Patents | Spain Applications | Sweden Patents | Sweden Applications | Switzerland Patents | Switzerland Applications | Ukraine Patents | Yugoslavia/Serbia and Montenegro Patents | Yugoslavia/Serbia and Montenegro Applications | Armenia Patents | Australia Patents | Australia Applications | Cyprus Patents | Gulf Cooperation Council Patents | Hong Kong Patents | India Patents | India Applications | Indonesia Patents | Indonesia Applications | Israel Applications | Jordan Patents | Jordan Applications | Kyrgyzstan Patents | Macao Applications | Malaysia Patents | Mongolia Patents | New Zealand Patents | Philippines Patents | Philippines Applications | Saudi Arabia Patents | Saudi Arabia Applications | Singapore Patents | Singapore Applications | Taiwan Patents | Taiwan Applications | Tajikistan Patents | Tajikistan Applications | Thailand Applications | Uzbekistan Patents | Vietnam Patents | Algeria Patents | ARIPO Patents | ARIPO Applications | Egypt Patents | Kenya Patents | Malawi Patents | Morocco Patents | Morocco Applications | OAPI Patents | South Africa Patents | Tunisia Applications | Zambia Patents | Zimbabwe Patents | EAPO Patents | EAPO Applications | Georgia Patents | Georgia Applications | Kazakhstan Patents | Kazakhstan Applications | Russia Patents | Russia Applications | Turkey Patents | Turkey Applications

1 2022-05-05 04:32 UTC | ~58,580,867 results from

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Text

1. A system for distributed execution of a sequence processing chain comprising: an interface adapted to set a measurement sequence for a plurality of measurement sites, each comprising a sequence runner; a sequencer repository adapted to be accessed locally from the plurality of measurement sites, and a sequence state manager adapted to receive measurement sequence states from at least one sequence runner and further adapted to distribute the measurement sequence states to other sequence runners via a network, wherein the measurement sequence states are associated with data and/or results through the sequence processing chain. 2. The system according to claim 1, wherein the sequence runners are adapted to compute resources available on measurement instruments in the plurality of measurement sites. 3. The system according to claim 1, wherein the sequence runners are further adapted to compute resources available on remote computing resources either located at the plurality of measurement sites or connected by the network. 4. The system according to claim 1, wherein the sequence runners on

the plurality of measurement sites are further adapted to determine whether a step of the sequence processing chain can be executed on non-measurement instruments and therefore conveys the measurement sequence states associated with the data and/or results to the sequence state manager for further execution of the step. 5. The system according to claim 4, wherein the sequence state manager is further adapted to determine resources available on the measurement instruments and on the remote computing resources and thereby distributes the measurement sequence states associated with the data and/or results to other sequence runners for further execution of the step. 6. The system according to claim 5, wherein the sequence state manager is further adapted to classify the steps of the sequence processing chain based on their dependency in order to execute the steps either in parallel or in sequence. 7. The system according to claim 6, wherein the sequence state manager is further adapted to allocate different resources for parallel execution of the steps and to allocate subsequent resources for sequential execution of the steps. 8. The system according to claim 5, wherein the sequence state manager is further adapted to determine the steps of the sequence processing chain that are required to be performed by the measurement instruments. 9. The system according to claim 8, wherein the remote computing resources are on a cloud computing platform and wherein the sequence state manager is further adapted to allocate the steps of the sequence processing chain that are not required to be performed by the measurement instruments to the cloud computing platform. 10. The system according to claim 9, wherein the data and/or results associated with the measurement sequence states are stored in the database of the cloud computing platform. 11. The system according to claim 2, wherein the measurement instruments are on a production line with devices under test, where the devices under test are respectively associated with an identification. 12. A method for distributed execution of a sequence processing chain comprising the steps of: setting a measurement sequence for a plurality of measurement sites, each comprising a sequence runner; receiving measurement sequence states from at least one sequence runner, and distributing the measurement sequence states to other sequence runners via a network, wherein the measurement sequence states are associated with data and/or results through the sequence processing chain. 13. The method according to claim 12, wherein the method further comprises the step of computing resources available on measurement instruments in the plurality of measurement sites. 14. The method according to claim 12, wherein the method further comprises the step of computing resources available on remote computers either located at the plurality of measurement sites or connected by the network. 15. The method according to claim 12, wherein the method further comprises the steps of: determining whether a step of the sequence processing chain can be executed on non-measurement instruments, and determining resources available on the measurement instruments and on the remote computers in order to distribute the measurement sequence states associated with the data and/or results to other sequence runners for further execution of the step. 16. The method according to claim 15, wherein the method further comprises the step of classifying the steps of the sequence processing chain based on their dependency in order to execute the steps either in parallel or in sequence. 17. The method according to claim 16, wherein the method further comprises the steps of: allocating different resources for parallel execution of the steps, and allocating subsequent resources for sequential execution of the steps. 18. The method according to claim 13, wherein the method further comprises the step of determining the steps of the sequence processing chain that are required to be performed by the measurement instruments. 19. The method according to claim 15, wherein the remote computing resources are on a cloud computing platform and wherein the method further comprises the step of allocating the steps of the sequence processing chain that are not required to be performed by the measurement instruments to the cloud computing platform. 20. The method according to claim 19, wherein the method further comprises the step of storing the data and/or results associated with the measurement sequence states in the database of the cloud computing platform.

Concept Modifiers: None

Filters: None

2 2022-05-05 04:33 UTC | ~17,620,602 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 1](#)

Concept Modifiers (1):

[1. More Like Document:](#)  US20210157637 System and method for distributed execution of a sequence processing chain [\(view\)](#)

Filters: None

3 2022-05-05 04:33 UTC | ~15,895,460 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: [Same as Search 1](#)

Concept Modifiers: [Same as Search 2](#)

Filters (1):

[Earliest Priority Date:](#) To 2019-11-22

4 2022-05-05 04:34 UTC | ~132,506 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: [Same as Search 1](#)

Concept Modifiers: [Same as Search 2](#)

Filters (2):

[Earliest Priority Date:](#) To 2019-11-22

[Classic Syntax limited to:](#) decentraliz*



Content: Patent Publications (2)

[US Patents](#) | [US Applications](#)

5 2022-05-05 04:34 UTC | ~37,444 results from 

De-Dup: Simple Family Numbers | **Cut-off:** None | **Sort:** Relevance

Main Concept: Text

1. A system for distributed execution of a sequence processing chain comprising: an interface adapted to set a measurement sequence for a plurality of measurement sites, each comprising a sequence runner; a sequencer repository adapted to be accessed locally from the plurality of measurement sites, and a sequence state manager adapted to receive measurement sequence states from at least one sequence runner and further adapted to distribute the measurement sequence states to other sequence runners via a network, wherein the measurement sequence states are associated with data and/or results through the sequence processing chain. 2. The system according to claim 1, wherein the sequence runners are adapted to compute resources available on measurement instruments in the plurality of measurement sites. 3. The system according to claim 1, wherein the sequence runners are further adapted to compute resources available on remote computing resources either located at the plurality of measurement sites or connected by the network. 4. The system according to claim 1, wherein the sequence runners on the plurality of measurement sites are further adapted to determine whether a step of the sequence processing chain

can be executed on non-measurement instruments and therefore conveys the measurement sequence states associated with the data and/or results to the sequence state manager for further execution of the step. 5. The system according to claim 4, wherein the sequence state manager is further adapted to determine resources available on the measurement instruments and on the remote computing resources and thereby distributes the measurement sequence states associated with the data and/or results to other sequence runners for further execution of the step. 6. The system according to claim 5, wherein the sequence state manager is further adapted to classify the steps of the sequence processing chain based on their dependency in order to execute the steps either in parallel or in sequence. 7. The system according to claim 6, wherein the sequence state manager is further adapted to allocate different resources for parallel execution of the steps and to allocate subsequent resources for sequential execution of the steps. 8. The system according to claim 5, wherein the sequence state manager is further adapted to determine the steps of the sequence processing chain that are required to be performed by the measurement instruments. 9. The system according to claim 8, wherein the remote computing resources are on a cloud computing platform and wherein the sequence state manager is further adapted to allocate the steps of the sequence processing chain that are not required to be performed by the measurement instruments to the cloud computing platform. 10. The system according to claim 9, wherein the data and/or results associated with the measurement sequence states are stored in the database of the cloud computing platform. 11. The system according to claim 2, wherein the measurement instruments are on a production line with devices under test, where the devices under test are respectively associated with an identification. 12. A method for distributed execution of a sequence processing chain comprising the steps of: setting a measurement sequence for a plurality of measurement sites, each comprising a sequence runner; receiving measurement sequence states from at least one sequence runner, and distributing the measurement sequence states to other sequence runners via a network, wherein the measurement sequence states are associated with data and/or results through the sequence processing chain. 13. The method according to claim 12, wherein the method further comprises the step of computing resources available on measurement instruments in the plurality of measurement sites. 14. The method according to claim 12, wherein the method further comprises the step of computing resources available on remote computers either located at the plurality of measurement sites or connected by the network. 15. The method according to claim 12, wherein the method further comprises the steps of: determining whether a step of the sequence processing chain can be executed on non-measurement instruments, and determining resources available on the measurement instruments and on the remote computers in order to distribute the measurement sequence states associated with the data and/or results to other sequence runners for further execution of the step. 16. The method according to claim 15, wherein the method further comprises the step of classifying the steps of the sequence processing chain based on their dependency in order to execute the steps either in parallel or in sequence. 17. The method according to claim 16, wherein the method further comprises the steps of: allocating different resources for parallel execution of the steps, and allocating subsequent resources for sequential execution of the steps. 18. The method according to claim 13, wherein the method further comprises the step of determining the steps of the sequence processing chain that are required to be performed by the measurement instruments. 19. The method according to claim 15, wherein the remote computing resources are on a cloud computing platform and wherein the method further comprises the step of allocating the steps of the sequence processing chain that are not required to be performed by the measurement instruments to the cloud computing platform. 20. The method according to claim 19, wherein the method further comprises the step of storing the data and/or results associated with the measurement sequence states in the database of the cloud computing platform.

Concept Modifiers (1):

[1. More Like Document:](#)  [US20210157637](#) System and method for distributed execution of a sequence processing chain [\(view\)](#)

Filters (2):

[Earliest Priority Date:](#) To 2019-11-22

[Classic Syntax limited to:](#) decentraliz*

6 2022-05-05 04:34 UTC | ~12,474 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (3):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

7 2022-05-05 04:36 UTC | ~1031 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (4):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

[Classic Syntax limited to: clon*](#)

8 2022-05-05 04:36 UTC | ~153 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (4):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

[Classic Syntax limited to: clon* with \(sequenc* or repository\)](#)

9 2022-05-05 04:41 UTC | 4 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (5):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

[Classic Syntax limited to: clon* with \(sequenc* or repository\)](#)

[CPC Codes limited to: G06F9/4881; G06F16/2379; G01R31/31718; G01R31/31722; G05B19/406; G06F9/5027; G05B2219/33297; G06F2209/503; G06Q10/06; G06F9/5072; G06N20/00](#)

10 2022-05-05 04:41 UTC | 3 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (6):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

[Classic Syntax limited to: clon* with \(sequenc* or repository\)](#)

[CPC Codes limited to: G06F9/4881; G06F16/2379; G01R31/31718; G01R31/31722; G05B19/406; G06F9/5027; G05B2219/33297; G06F2209/503; G06Q10/06; G06F9/5072; G06N20/00](#)

[Classic Syntax limited to: measur*](#)

11 2022-05-05 04:45 UTC | 0 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (5):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: chain](#)

[Classic Syntax limited to: clon* with \(sequenc* or repository\)](#)

[Classic Syntax limited to: "measurement site"](#)

12 2022-05-05 04:45 UTC | 0 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (4):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: clon* with \(sequenc* or repository\)](#)

[Classic Syntax limited to: "measurement site"](#)

13 2022-05-05 04:45 UTC | 0 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (4):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: clon*](#)

[Classic Syntax limited to: "measurement site"](#)

14 2022-05-05 04:45 UTC | ~559 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (4):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: clon*](#)

[Classic Syntax limited to: measurement](#)

15 2022-05-05 04:46 UTC | 8 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (5):

[Earliest Priority Date: To 2019-11-22](#)

[Classic Syntax limited to: decentraliz*](#)

[Classic Syntax limited to: clon*](#)

[Classic Syntax limited to: measurement](#)

[Classic Syntax limited to: repository same clon*](#)

16 2022-05-05 04:48 UTC | 6 results from 

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 5](#)

Concept Modifiers: [Same as Search 5](#)

Filters (6):

Earliest Priority Date: To 2019-11-22

Classic Syntax limited to: decentraliz*

Classic Syntax limited to: clon*

Classic Syntax limited to: measurement

Classic Syntax limited to: repository same clon*

Classic Syntax limited to: intermediate



Content: Patent Publications (166)

US Patents | US Designs | US Applications | EPO Patents | EPO Applications | China Patents | China Applications | Japan Patents | Japan Applications | Korea Patents | Korea Applications | WIPO Applications | Argentina Patents | Argentina Applications | Brazil Patents | Brazil Applications | Canada Patents | Canada Applications | Chile Patents | Chile Applications | Colombia Applications | Costa Rica Applications | Cuba Patents | Cuba Applications | Dominican Republic Applications | Ecuador Patents | Ecuador Applications | El Salvador Applications | Guatemala Applications | Honduras Applications | Mexico Patents | Mexico Applications | Nicaragua Patents | Panama Applications | Peru Applications | Trinidad & Tobago Patents | Uruguay Applications | Austria Patents | Austria Applications | Belarus Patents | Belgium Patents | Belgium Applications | Bosnia & Herzegovina Patents | Bosnia & Herzegovina Applications | Bulgaria Patents | Bulgaria Applications | Croatia Patents | Croatia Applications | Czech Republic Patents | Czech Republic Applications | Czechoslovakia Patents | Czechoslovakia Applications | Denmark Patents | Denmark Applications | Estonia Patents | Estonia Applications | EUIPO Patents | Finland Patents | Finland Applications | France Patents | France Applications | Germany Patents | Germany Applications | Great Britain Patents | Great Britain Applications | Greece Patents | Greece Applications | Hungary Patents | Hungary Applications | Iceland Patents | Iceland Applications | Ireland Patents | Ireland Applications | Italy Patents | Italy Applications | Latvia Patents | Latvia Applications | Lithuania Patents | Lithuania Applications | Luxembourg Patents | Malta Patents | Monaco Patents | Montenegro Patents | Montenegro Applications | Netherlands Patents | Netherlands Applications | Norway Patents | Norway Applications | Poland Patents | Poland Applications | Portugal Patents | Portugal Applications | Republic of Moldova Patents | Republic of Moldova Applications | Romania Patents | Romania Applications | San Marino Patents | San Marino Applications | Serbia Patents | Serbia Applications | Slovakia Patents | Slovakia Applications | Slovenia Patents | Spain Patents | Spain Applications | Sweden Patents | Sweden Applications | Switzerland Patents | Switzerland Applications | Ukraine Patents | Yugoslavia/Serbia and Montenegro Patents | Yugoslavia/Serbia and Montenegro Applications | Armenia Patents | Australia Patents | Australia Applications | Cyprus Patents | Gulf Cooperation Council Patents | Hong Kong Patents | India Patents | India Applications | Indonesia Patents | Indonesia Applications | Israel Applications | Jordan Patents | Jordan Applications | Kyrgyzstan Patents | Macao Applications | Malaysia Patents | Mongolia Patents | New Zealand Patents | Philippines Patents | Philippines Applications | Saudi Arabia Patents | Saudi Arabia Applications | Singapore Patents | Singapore Applications | Taiwan Patents | Taiwan Applications | Tajikistan Patents | Tajikistan Applications | Thailand Applications | Uzbekistan Patents | Vietnam Patents | Algeria Patents | ARIPO Patents | ARIPO Applications | Egypt Patents | Kenya Patents | Malawi Patents | Morocco Patents | Morocco Applications | OAPI Patents | South Africa Patents | Tunisia Applications | Zambia Patents | Zimbabwe Patents | EAPO Patents | EAPO Applications | Georgia Patents | Georgia Applications | Kazakhstan Patents | Kazakhstan Applications | Russia Patents | Russia Applications | Turkey Patents | Turkey Applications

17 2022-05-05 20:07 UTC | ~16,007,094 results from

De-Dup: Simple Family Numbers | Cut-off: None | Sort: Relevance

Main Concept: Text

1. A method of dynamically executing batches of requests on one or more execution engines running a machine-

learning transformer model, comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 2. The method of claim 1, further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 3. The method of claim 2, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 4. The method of claim 1, wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request. 5. The method of claim 4, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 6. The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 7. The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 8. The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 9. The method of claim 1, wherein each token in the sequence of input tokens represents a text unit. 10. A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 11. The non-transitory computer-readable storage medium of claim 10, the operations further comprising: responsive to determining that a request in the first

batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 12. The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 13. The non-transitory computer-readable storage medium of claim 10, wherein the second set of inputs include the sequence of input tokens for the new request and at least one output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request. 14. The non-transitory computer-readable storage medium of claim 13, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 15. The non-transitory computer-readable storage medium of claim 14, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 16. The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 17. The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 18. The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

Concept Modifiers: None

Filters: None

Content: Non-Patent Literature (23)

IEEE Xplore Publications: IEEE Periodicals | IEEE Xplore Publications: IEEE Conferences | IEEE Xplore Publications: IEEE Standards | IEEE Xplore Publications: IEEE Early Access | IEEE Xplore Publications: SMPTE Periodicals | IEEE Xplore Publications: SMPTE Conferences | IEEE Xplore Publications: SMPTE Standards | IEEE Xplore Publications: MIT Press eBooks | IEEE Xplore Publications: Wiley-IEEE eBooks | IEEE Xplore Publications: IBM Periodicals | IEEE Xplore Publications: URSI Periodicals | IEEE Xplore Publications: VDE Conferences | IEEE Xplore Publications: Periodicals from China | IP.com Prior Art Database: The IP.com Journal | IP.com Prior Art Database: Internet Society RFC | IP.com Prior Art Database: IBM TDB Archive | IP.com Prior Art Database: Legacy Journals | IP.com Prior Art Database: Software Patent Institute | OnePetro.org: Periodicals at OnePetro.org | OnePetro.org: Conferences at OnePetro.org | Other Literature: IBM Redbooks | Other Literature: PubMed Central | Other Literature: arXiv.org

18 2022-05-05 20:07 UTC | 2,142,326 results from 

De-Dup: None | **Cut-off:** None | **Sort:** Relevance

Main Concept: Text

1. A method of dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning

transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 2. The method of claim 1, further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one or more requests and the new request. 3. The method of claim 2, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 4. The method of claim 1, wherein the second set of inputs include the sequence of input tokens for the new request and an output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the output token for the at least one request. 5. The method of claim 4, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 6. The method of claim 5, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 7. The method of claim 1, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 8. The method of claim 1, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 9. The method of claim 1, wherein each token in the sequence of input tokens represents a text unit. 10. A non-transitory computer-readable storage medium storing computer program instructions executable to perform operations for dynamically executing batches of requests on one or more execution engines running a machine-learning transformer model, the operations comprising: receiving, by a serving system, one or more requests for execution, the serving system including a scheduler and one or more execution engines each coupled to access a machine-learning transformer model including at least a set of decoders; scheduling, by the scheduler, a batch of requests including the one or more requests for execution on an execution engine; generating, by the execution engine, a first set of output tokens by applying the transformer model to a first set of inputs for the batch of requests, wherein applying the transformer model comprises applying at least one batch operation to one or more input tensors associated with the batch of requests; receiving, by a request processor, a new request from a client device, the new request including a sequence of input tokens; scheduling, by the scheduler, a second batch of requests additionally including the new request for execution on the execution engine, the second batch of requests scheduled responsive to determining that the execution engine has memory available to execute the second batch of requests; and generating, by the execution engine, a second set of output tokens by applying the transformer model to a second set of inputs for the second batch of requests including the sequence of input tokens for the new request. 11. The non-transitory computer-readable storage medium of claim 10, the operations further comprising: responsive to determining that a request in the first batch of requests has been completed, providing output tokens generated for the completed request to a client device as a response to the request, and wherein the second batch of requests includes the remaining requests from the one

or more requests and the new request. 12. The non-transitory computer-readable storage medium of claim 11, wherein the request is associated with a memory cache in the execution engine dedicated for storing an internal state for the request, and responsive to determining that the request has been completed, freeing the dedicated memory cache for the request in the execution engine. 13. The non-transitory computer-readable storage medium of claim 10, wherein the second set of inputs include the sequence of input tokens for the new request and at least one output token from the first set of output tokens for at least one request, and wherein a length of the sequence of input tokens for the new request is different from a length of the at least one output token for the at least one request. 14. The non-transitory computer-readable storage medium of claim 13, wherein the execution engine includes a cache memory for maintaining a key cache tensor for storing keys and a value cache tensor for storing values for the at least one request, and wherein after scheduling the second batch of requests, allocating, by the execution engine, a new cache memory dedicated to maintaining a key cache tensor and a value cache tensor for the new request. 15. The non-transitory computer-readable storage medium of claim 14, wherein after generating the second set of output tokens, a length of the key cache tensor for the at least one request is different from a length of a key cache tensor for the new request, and a length of the value cache tensor for the at least one request is different from a length of a value cache tensor for the new request. 16. The non-transitory computer-readable storage medium of claim 10, after receiving the new request from the client device, determining, by the scheduler, that there is insufficient memory available to execute the second batch of requests on a second execution engine different from the execution engine, and responsive to the determination for the second execution engine, determining whether the execution engine has the memory available to execute the second batch of requests. 17. The non-transitory computer-readable storage medium of claim 10, wherein the execution engine is configured as a graphics processing unit (GPU) or a tensor processing unit (TPU). 18. The non-transitory computer-readable storage medium of claim 10, wherein each token in the sequence of input tokens represents a text unit.

Concept Modifiers: None

Filters: None

19 2022-05-05 20:08 UTC | 2,046,771 results from 

De-Dup: None | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 18](#)

Concept Modifiers: None

Filters (1):

[Publication Date](#): To 2021-12-03

20 2022-05-05 20:08 UTC | 143,885 results from 

De-Dup: None | Cut-off: None | Sort: Relevance

Main Concept: [Same as Search 18](#)

Concept Modifiers: None

Filters (2):

[Publication Date](#): To 2021-12-03

[Classic Syntax limited to](#): transform* and batch*

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	1	of	2	Attorney Docket Number	39192-50678/US

U.S. PATENT DOCUMENTS				
		Document No.		
Examiner Initials*	Cite No. ¹	Number – Kind Code ² (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document
	A1	US-2021/0034335 A1	02-04-2021	Svyalkovskly et al.
	A2	US-2021/0357210 A1	11-18-2021	Clement et al.
	A3	US-2022/0066747 A1	03-03-2022	Drain et al.
	A4	US-2022/0067513 A1	03-03-2022	Stevens et al.

FOREIGN PATENT DOCUMENTS					
		Foreign Patent Document			
Examiner Initials*	Cite No. ¹	Country Code ³ – Number ⁴ Kind Code ⁵ (if known)	Publication Date MM-DD-YYYY	Name of Patentee or Applicant of Cited Document	T ⁶
	B1	N/A			

OTHER REFERENCES – NON-PATENT LITERATURE DOCUMENTS				
Examiner Initials*	Cite No. ¹	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published		T ⁶
	C1	DOSHI, Ketan, "Transformers Explained Visually (Part 1): Overview of Functionality", Dec.13, 2020, <towardsdatascience.com> (Year: 2020), 16 pages.		
	C2	DOSHI, Ketan, "Transformers Explained Visually (Part 2): How it works, step-by-step", Jan.2, 2021, <towardsdatascience.com> (Year: 2021), 23 pages.		
	C3	DOSHI, Ketan, "Transformers Explained Visually (Part 3): Multi-head Attention deep dive", Jan.16, 2021, <towardsdatascience.com> (Year: 2021), 20 pages.		
	C4	DOSHI, Ketan, "Transformers Explained Visually (Part 4): Not Just How, but Why They Work So Well", June 2, 2021, <towardsdatascience.com> (Year: 2021), 17 pages.		
	C5	VASWANI et al., "Attention Is All You Need", Dec. 6, 2017, arXiv, <https://arxiv.org/abs/1706.03762> (Year: 2017), pages 1-15.		

EXAMINER SIGNATURE	
Examiner Signature /KENNETH TANG/	
Date Considered 05/05/2022	
*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.	
¹ Applicant's unique citation designation number (option). ² See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ³ Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ⁴ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document.	
⁵ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁶ Applicant is to place a check mark here if English language translation is attached.	

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)				Complete if Known	
				Application Number	17/542,193
				Filing Date	December 3, 2021
				First Named Inventor	Gyeongin Yu
				Applicant	FriendliAI Inc.
				Art Unit	2199
				Confirmation Number	1026
				Examiner Name	TANG, KENNETH
Sheet	2	of	2	Attorney Docket Number	39192-50678/US

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☒ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

☐ After payment of the issue fee. Applicant requests that the information contained in this Information Disclosure Statement be placed in the file according to 37 CFR § 1.97(i), although the information may not be considered by the USPTO.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Jae Yeon Baek/	Date (YYYY-MM-DD)	2022-04-12
Name/Print	Jae Yeon Baek	Registration Number	78,258

POWER OF ATTORNEY TO PROSECUTE APPLICATIONS BEFORE THE USPTO

I hereby revoke all previous powers of attorney given in the application identified in the attached statement under 37 CFR 3.73(c).

I hereby appoint:

☒ Practitioners associated with Customer Number: **00758**

OR

☐ Practitioner(s) named below (if more than ten patent practitioners are to be named, then a customer number must be used):

Name	Registration Number

Name	Registration Number

As attorney(s) or agent(s) to represent the undersigned before the United States Patent and Trademark Office (USPTO) in connection with any and all patent applications assigned only to the undersigned according to the USPTO assignment records or assignment documents attached to this form in accordance with 37 CFR 3.73(c).

Please change the correspondence address for the application identified in the attached statement under 37 CFR 3.73(c) to:

☒ The address associated with Customer Number: **00758**

OR

Firm or individual name		
Address		
City	State	Zip
Country		
Telephone	Email	

Assignee name and address: **FriendliAI Inc., Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea**

A copy of this form, together with a statement under 37 CFR 3.73(c) (Form PTO/AIA/96 or equivalent) is required to be filed in each application in which this form is used. The statement under 37 CFR 3.73(c) may be completed by one of the practitioners appointed in this form, and must identify the application in which this Power of Attorney is to be filed.

SIGNATURE of Assignee of Record

The individual whose signature and title is supplied below is authorized to act on behalf of the assignee.

Signature <i>Byung-Gon Chun</i>	Date <i>May 18, 2022</i>
Name <i>Byung-Gon Chun</i>	Telephone <i>82-2-889-8020</i>
Title <i>CEO</i>	

This collection of information is required by 37 CFR 1.31, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public, which is to update (and by the USPTO to process) the file of a patent or reexamination proceeding. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 18 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	45756026
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Won Song
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	19-MAY-2022
Filing Date:	03-DEC-2021
Time Stamp:	15:14:56
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no
------------------------	----

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Request under 37CFR1.46(c) to correct, update, change applicant	50678_Request_to_Change_Appl applicant.pdf	195012	no	2
			e9b966e48be1684378b4bd59d230974019390891		

Warnings:

239 of 265

Information:					
2	Application Data Sheet	50678_Supplemental_Application_Data_Sheet.pdf	408007 9cb0cc7871329e27188d113914536af9cb9a831c	no	9
Warnings:					
Information:					
This is not an USPTO supplied ADS fillable form					
3	Assignee showing of ownership per 37 CFR 3.73	50678_373c_Statement.pdf	113543 ea68d63761a9cd772632bd16486fda27237729c	no	3
Warnings:					
Information:					
4	Power of Attorney	50678_Power_of_Attorney.pdf	1227497 97b95baad3e7168097d42269358549aae09a3337	no	1
Warnings:					
Information:					
Total Files Size (in bytes):			1944059		
This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503. <u>New Applications Under 35 U.S.C. 111</u> If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application. <u>National Stage of an International Application under 35 U.S.C. 371</u> If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course. <u>New International Application Filed with the USPTO as a Receiving Office</u> If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.					

Doc Code: R46C.REQ

Document Description: Request under 37 CFR 1.46(c) to correct, update or change the applicant.

PTO/AIA/41 (04-15)

Approved for use through 11/30/2020. OMB 0651-0031

U.S. Patent and Trademark Office, U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

**REQUEST TO CORRECT OR UPDATE
THE NAME OF THE APPLICANT UNDER
37 CFR 1.46(c)(1), OR CHANGE THE
APPLICANT UNDER 37 CFR 1.46(c)(2)
(FOR USE ONLY IN APPLICATIONS FILED ON OR AFTER
SEPTEMBER 16, 2012)**

Application Number	17/542,193
Filing Date	December 3, 2021
First Named Inventor	Gyeongin Yu
Art Unit	2199
Examiner Name	Kenneth Tang
Practitioner Docket Number	39192-50678/US

**To: Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450**

Applicant hereby **requests that the name of the applicant be corrected or updated under 37 CFR 1.46(c)(1), or that the applicant be changed under 37 CFR 1.46(c)(2),** in the above-identified application. **Requests under 37 CFR 1.46(c)(1) or (c)(2) cannot be submitted after payment of the issue fee or if the application has been patented.**

Please check the applicable box(es) below.

☐ 1. This request is to **correct or update the name of the applicant (under 37 CFR 1.46(c)(1))** and includes:

☐ An application data sheet (ADS) in accordance with 37 CFR 1.76(c) with the corrected or updated information shown with markings (e.g., underlining for insertions, strikethrough for deletions). A Corrected Web-based ADS may be used.

Note: Requests under 37 CFR 1.46(c)(1) may be filed to correct typographical errors in the name of the § 1.46 applicant, or for updating the name of the § 1.46 applicant (i.e., where there is no change in the applicant itself but just in the applicant's name). See the Manual of Patent Examining Procedure (MPEP) section 605.01.

☒ 2. This request is to **change the applicant (under 37 CFR 1.46(c)(2))** and includes:

☒ An application data sheet (ADS) in accordance with 37 CFR 1.76(c) that identifies the changes with proper markings (underlining for insertions and strikethrough for deletions). A Corrected Web-based ADS may be used.

☒ A Statement Under 37 CFR 3.73(c) (Form PTO/AIA/96 or equivalent). See MPEP 325.

I am the

☐ applicant*

☒ attorney or agent of record

Registration number 59,070

☐ attorney or agent acting under 37 CFR 1.34

Registration number _____

Signature /Jae Won Song/

Typed or printed name Jae Won Song

Date May 19, 2022

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. *Juristic entities must be represented by a patent practitioner (See 37 CFR 1.31, applicable to any paper filed on or after September 16, 2012 that is presented on behalf of a juristic entity, regardless of application filing date). Submit multiple forms if more than one signature is required, see below**.

☒ ** Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.36. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND**

TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	17/542,193
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76. This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.			

Secrecy Order 37 CFR 5.2:

☐ Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)
--

Inventor Information:

Inventor 1					Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Gyeongin		Yu		
Residence Information (Select One) ☐ US Residency ☒ Non US Residency ☐ Active US Military Service					
City	Seoul	Country of Residence ⁱ	KR		
Mailing Address of Inventor:					
Address 1		FriendliAI Inc.			
Address 2		Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province			
Postal Code		08826	Country ⁱ	KR	
Inventor 2					Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Geon-Woo		Kim		
Residence Information (Select One) ☐ US Residency ☒ Non US Residency ☐ Active US Military Service					
City	Seoul	Country of Residence ⁱ	KR		
Mailing Address of Inventor:					
Address 1		FriendliAI Inc.			
Address 2		Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province			
Postal Code		08826	Country ⁱ	KR	
Inventor 3					Remove
Legal Name					

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	17/542,193
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Prefix	Given Name	Middle Name	Family Name	Suffix
	Joo Seong		Jeong	
Residence Information (Select One) ☐ US Residency ☒ Non US Residency ☐ Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:				
Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	
Inventor 4				Remove
Legal Name				

Prefix	Given Name	Middle Name	Family Name	Suffix
	Soojeong		Kim	
Residence Information (Select One) ☐ US Residency ☒ Non US Residency ☐ Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:				
Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	
Inventor 5				Remove
Legal Name				

Prefix	Given Name	Middle Name	Family Name	Suffix
	Byung-Gon		Chun	
Residence Information (Select One) ☐ US Residency ☒ Non US Residency ☐ Active US Military Service				
City	Seoul	Country of Residence ⁱ	KR	

Mailing Address of Inventor:				
Address 1	FriendliAI Inc.			
Address 2	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu			
City	Seoul	State/Province		
Postal Code	08826	Country ⁱ	KR	

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	<u>17/542,193</u>
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button.

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

☐ An Address is being provided for the correspondence Information of this application.

Customer Number	00758		
Email Address	ptoc@fenwick.com	Add Email	Remove Email

Application Information:

Title of the Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		
Attorney Docket Number	39192-50678/US	Small Entity Status Claimed	☒
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	12	Suggested Figure for Publication (if any)	

Filing By Reference:

Only complete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32). Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

SUPPLEMENTAL		Attorney Docket Number		39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number		17/542,193
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks			
Please Select One:	☒ Customer Number	☐ US Patent Practitioner	☐ Limited Recognition (37 CFR 11.9)	
Customer Number	00758			

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, 365(c), or 386(c) or indicate National Stage entry from a PCT application. Providing benefit claim information in the Application Data Sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78. When referring to the current application, please leave the "Application Number" field blank.

Prior Application Status			Remove
Application Number	Continuity Type	Prior Application Number	Filing or 371(c) Date (YYYY-MM-DD)
Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the Add button.			

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55. When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX)ⁱ the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(i)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

Application Number	Country ⁱ	Filing Date (YYYY-MM-DD)	Remove	Access Code ⁱ (if applicable)
Additional Foreign Priority Data may be generated within this form by selecting the Add button.				

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

☐ This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013. NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.

SUPPLEMENTAL Application Data Sheet 37 CFR 1.76	Attorney Docket Number	39192-50678/US
	Application Number	<u>17/542,193</u>
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks	

Authorization or Opt-Out of Authorization to Permit Access:

When this Application Data Sheet is properly signed and filed with the application, applicant has provided written authority to permit a participating foreign intellectual property (IP) office access to the instant application-as-filed (see paragraph A in subsection 1 below) and the European Patent Office (EPO) access to any search results from the instant application (see paragraph B in subsection 1 below).

Should applicant choose not to provide an authorization identified in subsection 1 below, applicant **must opt-out** of the authorization by checking the corresponding box A or B or both in subsection 2 below.

NOTE: This section of the Application Data Sheet is **ONLY** reviewed and processed with the **INITIAL** filing of an application. After the initial filing of an application, an Application Data Sheet cannot be used to provide or rescind authorization for access by a foreign IP office(s). Instead, Form PTO/SB/39 or PTO/SB/69 must be used as appropriate.

1. Authorization to Permit Access by a Foreign Intellectual Property Office(s)

A. Priority Document Exchange (PDX) - Unless box A in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the World Intellectual Property Organization (WIPO), and any other foreign intellectual property office participating with the USPTO in a bilateral or multilateral priority document exchange agreement in which a foreign application claiming priority to the instant patent application is filed, access to: (1) the instant patent application-as-filed and its related bibliographic data, (2) any foreign or domestic application to which priority or benefit is claimed by the instant application and its related bibliographic data, and (3) the date of filing of this Authorization. See 37 CFR 1.14(h)(1).

B. Search Results from U.S. Application to EPO - Unless box B in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the EPO access to the bibliographic data and search results from the instant patent application when a European patent application claiming priority to the instant patent application is filed. See 37 CFR 1.14(h)(2).

The applicant is reminded that the EPO's Rule 141(1) EPC (European Patent Convention) requires applicants to submit a copy of search results from the instant application without delay in a European patent application that claims priority to the instant application.

2. Opt-Out of Authorizations to Permit Access by a Foreign Intellectual Property Office(s)

☐ A. Applicant **DOES NOT** authorize the USPTO to permit a participating foreign IP office access to the instant application-as-filed. If this box is checked, the USPTO will not be providing a participating foreign IP office with any documents and information identified in subsection 1A above.

☐ B. Applicant **DOES NOT** authorize the USPTO to transmit to the EPO any search results from the instant patent application. If this box is checked, the USPTO will not be providing the EPO with search results from the instant application.

NOTE: Once the application has published or is otherwise publicly available, the USPTO may provide access to the application in accordance with 37 CFR 1.14.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	<u>17/542,193</u>
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.			
Applicant 1			
If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.			
Clear			
☒ Assignee		☐ Legal Representative under 35 U.S.C. 117	
☐ Person to whom the inventor is obligated to assign.		☐ Joint Inventor	
☐ Person who shows sufficient proprietary interest			
If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:			
Name of the Deceased or Legally Incapacitated Inventor:			
If the Applicant is an Organization check here. ☒			
Organization Name	FriendliAI Inc.		
Mailing Address Information For Applicant:			
Address 1	Rm. 514, Bldg. 138, Gwanak-ro, Gwanak-gu		
Address 2			
City	Seoul	State/Province	
Country	KR	Postal Code	08826
Phone Number		Fax Number	
Email Address			
Additional Applicant Data may be generated within this form by selecting the Add button.			

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	17/542,193
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Applicant 2			
If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section. Clear			
☒ Assignee		☐ Legal Representative under 35 U.S.C. 117	☐ Joint Inventor
☐ Person to whom the inventor is obligated to assign.		☐ Person who shows sufficient proprietary interest	
If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:			
Name of the Deceased or Legally Incapacitated Inventor:			
If the Applicant is an Organization check here. ☒			
Organization Name	~~Seoul National University R&DB Foundation~~		
Mailing Address Information For Applicant:			
Address 1	~~1, Gwanak-ro, Gwanak-gu~~		
Address 2			
City	~~Seoul~~	State/Province	
Country	~~KR~~	Postal Code	~~00020~~
Phone Number		Fax Number	
Email Address			
Additional Applicant Data may be generated within this form by selecting the Add button.			

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Assignee 1
Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication. An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication.
If the Assignee or Non-Applicant Assignee is an Organization check here. ☐

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

SUPPLEMENTAL		Attorney Docket Number	39192-50678/US
Application Data Sheet 37 CFR 1.76		Application Number	<u>17/542,193</u>
Title of Invention	Dynamic Batching for Inference System for Transformer-Based Generation Tasks		

Prefix	Given Name	Middle Name	Family Name	Suffix

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1			
Address 2			
City		State/Province	
Country		Postal Code	
Phone Number		Fax Number	
Email Address			

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button.

Signature:

NOTE: This Application Data Sheet must be signed in accordance with 37 CFR 1.33(b). However, if this Application Data Sheet is submitted with the **INITIAL** filing of the application and either box A or B is not checked in subsection 2 of the "Authorization or Opt-Out of Authorization to Permit Access" section, then this form must also be signed in accordance with 37 CFR 1.14(c).

This Application Data Sheet **must** be signed by a patent practitioner if one or more of the applicants is a **juristic entity** (e.g., corporation or association). If the applicant is two or more joint inventors, this form must be signed by a patent practitioner, **all** joint inventors who are the applicant, or one or more joint inventor-applicants who have been given power of attorney (e.g., see USPTO Form PTO/AIA/81) on behalf of **all** joint inventor-applicants.

See 37 CFR 1.4(d) for the manner of making signatures and certifications.

Signature	/Jae Won Song/		Date (YYYY-MM-DD)	2022-05-19
First Name	Jae Won	Last Name	Song	Registration Number
				59,070

Additional Signature may be generated within this form by selecting the Add button.

This collection of information is required by 37 CFR 1.76. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 23 minutes to complete, including gathering, preparing, and submitting the completed application data sheet form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

- 1 The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
- 2 A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
- 3 A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
- 4 A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
- 5 A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
- 6 A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
- 7 A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
- 8 A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
- 9 A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

STATEMENT UNDER 37 CFR 3.73(c)Applicant/Patent Owner: FriendliAI Inc.Application No./Patent No.: 17/542,193Filed/Issue Date: December 3, 2021Titled: Dynamic Batching for Inference System for Transformer-Based Generation TasksFriendliAI Inc., a corporation

(Name of Assignee)

(Type of Assignee, e.g., corporation, partnership, university, government agency, etc.)

states that, for the patent application/patent identified above, it is (choose one of options 1, 2, 3 or 4 below):

1. ☒ The assignee of the entire right, title, and interest.
2. ☐ An assignee of less than the entire right, title, and interest (check applicable box):
- ☐ The extent (by percentage) of its ownership interest is _____%. Additional Statement(s) by the owners holding the balance of the interest must be submitted to account for 100% of the ownership interest.
- ☐ There are unspecified percentages of ownership. The other parties, including inventors, who together own the entire right, title and interest are:

Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

3. ☐ The assignee of an undivided interest in the entirety (a complete assignment from one of the joint inventors was made). The other parties, including inventors, who together own the entire right, title, and interest are:

Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

4. ☐ The recipient, via a court proceeding or the like (e.g., bankruptcy, probate), of an undivided interest in the entirety (a complete transfer of ownership interest was made). The certified document(s) showing the transfer is attached.

The interest identified in option 1, 2 or 3 above (not option 4) is evidenced by either (choose one of options A or B below):

- A. ☐ An assignment from the inventor(s) of the patent application/patent identified above. The assignment was recorded in the United States Patent and Trademark Office at Reel _____, Frame _____, or for which a copy thereof is attached.
- B. ☒ A chain of title from the inventor(s), of the patent application/patent identified above, to the current assignee as follows:

1. From: Gyeongin YU, Geon-Woo KIM, Joo Seong JEONG, Soojeong KIM, Byung-Gon CHUN To: FriendliAI Inc. and Seoul National University R&DB FoundationThe document was recorded in the United States Patent and Trademark Office at
Reel 058599, Frame 0509, or for which a copy thereof is attached.2. From: Seoul National University R&DB Foundation To: FriendliAI Inc.The document was recorded in the United States Patent and Trademark Office at
Reel 059935, Frame 0001, or for which a copy thereof is attached.

[Page 1 of 2]

This collection of information is required by 37 CFR 3.73(b). The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

STATEMENT UNDER 37 CFR 3.73(c)

3. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at
 Reel _____, Frame _____, or for which a copy thereof is attached.

4. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at
 Reel _____, Frame _____, or for which a copy thereof is attached.

5. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at
 Reel _____, Frame _____, or for which a copy thereof is attached.

6. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at
 Reel _____, Frame _____, or for which a copy thereof is attached.

☐ Additional documents in the chain of title are listed on a supplemental sheet(s).

☒ As required by 37 CFR 3.73(c)(1)(i), the documentary evidence of the chain of title from the original owner to the assignee was, or concurrently is being, submitted for recordation pursuant to 37 CFR 3.11.

[NOTE: A separate copy (i.e., a true copy of the original assignment document(s)) must be submitted to Assignment Division in accordance with 37 CFR Part 3, to record the assignment in the records of the USPTO. See MPEP 302.08]

The undersigned (whose title is supplied below) is authorized to act on behalf of the assignee.

/Jae Won Song/

5/19/2022

Signature

Date

Jae Won Song

59,070

Printed or Typed Name

Title or Registration Number

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY. DOCKET NO	TOT CLAIMS	IND CLAIMS
17/542,193	12/03/2021	2199	910	39192-50678/US	18	2

CONFIRMATION NO. 1026

CORRECTED FILING RECEIPT



000000013399837

758

FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

Date Mailed: 05/23/2022

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Gyeongin Yu, Seoul, KOREA, REPUBLIC OF;
Geon-Woo Kim, Seoul, KOREA, REPUBLIC OF;
Joo Seong Jeong, Seoul, KOREA, REPUBLIC OF;
Soojeong Kim, Seoul, KOREA, REPUBLIC OF;
Byung-Gon Chun, Seoul, KOREA, REPUBLIC OF;

Applicant(s)

FriendliAI Inc., Seoul, KOREA, REPUBLIC OF;

Power of Attorney: The patent practitioners associated with Customer Number 00758

Domestic Applications for which benefit is claimed - None.

A proper domestic benefit claim must be provided in an Application Data Sheet in order to constitute a claim for domestic benefit. See 37 CFR 1.76 and 1.78.

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see <http://www.uspto.gov> for more information.) - None.

Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 12/14/2021

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 17/542,193**

Projected Publication Date: 06/08/2023

Non-Publication Request: No

Early Publication Request: No

**** SMALL ENTITY ****

Title

Dynamic Batching for Inference System for Transformer-Based Generation Tasks

Preliminary Class

718

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at <http://www.uspto.gov/web/offices/pac/doc/general/index.html>.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, <http://www.stopfakes.gov>. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific

countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER
Title 35, United States Code, Section 184
Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop

technology, manufacture products, deliver services, and grow your business, visit <http://www.SelectUSA.gov> or call +1-202-482-6800.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US

CONFIRMATION NO. 1026

POA ACCEPTANCE LETTER



758
FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

Date Mailed: 05/23/2022

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 05/19/2022.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at (571) 272-4000 or (571) 272-4200 or 1-888-786-0101.

/ldvan/

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via EFS-Web.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

758 7590 05/11/2022
FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via EFS-Web or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	12/03/2021	Gyeongin Yu	39192-50678/US	1026

TITLE OF INVENTION: Dynamic Batching for Inference System for Transformer-Based Generation Tasks

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	\$600	\$0.00	\$0.00	\$600	08/11/2022

EXAMINER	ART UNIT	CLASS-SUBCLASS
TANG, KENNETH	2199	718-101000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

☐ Change of correspondence address (for Change of Correspondence Address form PTO/AIA/122 or PTO/SB/122) attached.

☐ "Fee Address" indication (or "Fee Address" Indication form PTO/AIA/47 or PTO/SB/47; Rev 03-02 or more recent) attached. Use of a Customer Number is required.

2. For printing on the patent front page, list

(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,

(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 Fenwick & West LLP

2

3

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE

(B) RESIDENCE: (CITY and STATE OR COUNTRY)

FriendliAI Inc.

Seoul, Republic of Korea

Please check the appropriate assignee category or categories (will not be printed on the patent): ☐ Individual ☒ Corporation or other private group entity ☐ Government

4a. Fees submitted: ☒ Issue Fee ☐ Publication Fee (if required) ☐ Advance Order - # of Copies _____

4b. Method of Payment: (Please first reapply any previously paid fee shown above)

☒ Electronic Payment via EFS-Web ☐ Enclosed check ☐ Non-electronic payment by credit card (Attach form PTO-2038)

☒ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. 192555

5. Change in Entity Status (from status indicated above)

☐ Applicant certifying micro entity status. See 37 CFR 1.29

☐ Applicant asserting small entity status. See 37 CFR 1.27

☐ Applicant changing to regular undiscounted fee status.

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.

NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.

NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature /Jae Won Song/

Date August 4, 2022

Typed or printed name Jae Won Song

Registration No. 59,070

Electronic Patent Application Fee Transmittal

Application Number:	17542193			
Filing Date:	03-Dec-2021			
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks			
First Named Inventor/Applicant Name:	Gyeongin Yu			
Filer:	Jae Won Song/Carol Bergeron			
Attorney Docket Number:	39192-50678/US			
Filed as Small Entity				
Filing Fees for Utility under 35 USC 111(a)				
Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
UTILITY APPL ISSUE FEE	2501	1	600	600

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Extension-of-Time:				
Miscellaneous:				
Total in USD (\$)				600

Electronic Acknowledgement Receipt

EFS ID:	46347315
Application Number:	17542193
International Application Number:	
Confirmation Number:	1026
Title of Invention:	Dynamic Batching for Inference System for Transformer-Based Generation Tasks
First Named Inventor/Applicant Name:	Gyeongin Yu
Customer Number:	758
Filer:	Jae Won Song
Filer Authorized By:	
Attorney Docket Number:	39192-50678/US
Receipt Date:	04-AUG-2022
Filing Date:	03-DEC-2021
Time Stamp:	21:24:38
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	\$600
RAM confirmation Number	E202284L25432784
Deposit Account	192555
Authorized User	Jae Won Song

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

37 CFR 1.16 (National application filing, search, and examination fees)

37 CFR 1.17 (Patent application and reexamination processing fees)

2039 of 205

37 CFR 1.19 (Document supply fees)
 37 CFR 1.20 (Post Issuance fees)
 37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Issue Fee Payment (PTO-85B)	50678_Issue_Fee_Transmittal.pdf	290073	no	1
			5d589e35cd972913fce1565e0b4e23f8d7ba145		

Warnings:

Information:

2	Fee Worksheet (SB06)	fee-info.pdf	37905	no	2
			ba429591df6c35c1425806ab9326debaa3718e9d		

Warnings:

Information:

Total Files Size (in bytes):	327978
-------------------------------------	--------

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.



APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/542,193	09/13/2022	11442775	39192-50678/US	1026

758 7590 08/24/2022

FENWICK & WEST LLP
SILICON VALLEY CENTER
801 CALIFORNIA STREET
MOUNTAIN VIEW, CA 94041

ISSUE NOTIFICATION

The projected patent number and issue date are specified above.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b) (application filed on or after May 29, 2000)

The Patent Term Adjustment is 0 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (<http://pair.uspto.gov>).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Patents Stakeholder Experience (OPSE), Stakeholder Support Division (SSD) at (571)-272-4200.

INVENTOR(s) (Please see PAIR WEB site <http://pair.uspto.gov> for additional inventors):

Gyeongin Yu, Seoul, KOREA, REPUBLIC OF;
Geon-Woo Kim, Seoul, KOREA, REPUBLIC OF;
Joo Seong Jeong, Seoul, KOREA, REPUBLIC OF;
Soojeong Kim, Seoul, KOREA, REPUBLIC OF;
Byung-Gon Chun, Seoul, KOREA, REPUBLIC OF;

APPLICANT(s) (Please see PAIR WEB site <http://pair.uspto.gov> for additional applicants):

FriendliAI Inc., Seoul, KOREA, REPUBLIC OF;

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The USA offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to encourage and facilitate business investment. To learn more about why the USA is the best country in the world to develop technology, manufacture products, and grow your business, visit SelectUSA.gov.