

UNITED STATES PATENT AND TRADEMARK OFFICE

---

BEFORE THE PATENT TRIAL AND APPEAL BOARD

---

CROWDSTRIKE, INC.,  
Petitioner

v.

SKYSONG INNOVATIONS, LLC,  
Patent Owner

---

*Inter Partes* Review Case No. IPR2025-01399  
U.S. Patent No. 11,775,831

**DECLARATION OF JUNE ANN MUNFORD**

## **Introduction and Qualifications**

1. My name is June Ann Munford. I am over the age of 18, have personal knowledge of the facts set forth herein, and am competent to testify to the same.

2. I earned a Master of Library and Information Science (MLIS) from the University of Wisconsin-Milwaukee in 2009. I have over ten years of experience in the library/information science field. Beginning in 2004, I have served in various positions in the public library sector including Assistant Librarian, Youth Services Librarian and Library Director. I have attached my Curriculum Vitae as Appendix CV.

3. During my career in the library profession, I have been responsible for materials acquisition for multiple libraries. In that position, I have cataloged, purchased and processed incoming library works. That includes purchasing materials directly from vendors, recording publishing data from the material in question, creating detailed material records for library catalogs and physically preparing that material for circulation. In addition to my experience in acquisitions, I was also responsible for analyzing large collections of library materials, tailoring library records for optimal catalog search performance and creating lending agreements between libraries during my time as a Library Director.

## **Library Cataloging and Bibliographic Practices**

4. I am fully familiar with the catalog record creation process in the library sector. In preparing a material for public availability, a library catalog record describing that material would be created. These records are typically written in Machine Readable Catalog (herein referred to as “MARC”) code and contain information such as a physical description of the material, metadata from the material’s publisher, and date of library acquisition. In particular, the 008 field of the MARC record is reserved for denoting the date of creation of the library record itself. As this typically occurs during the process of preparing materials for public access, it is my experience that an item’s MARC record indicates the date of an item’s public availability.

5. Typically, in creating a MARC record, a librarian would gather various bits of metadata such as book title, publisher and subject headings among others and assign each value to a relevant numerical field. For example, a book’s physical description is tracked in field 300 while title/attribution is tracked in field 245. The 008 field of the MARC record is reserved for denoting the creation of the library record itself. As this is the only date reflecting the inclusion of said materials within the library’s collection, it is my experience that an item’s 008 field accurately indicates the date of an item’s public availability.

6. Libraries typically catalog broader publications, such as conference proceedings, under the title of the proceedings rather than under the individual papers included within them. Accordingly, under standard library practices reflected in MARC records, any paper published as part of a proceedings volume would have been available to the public once that proceedings volume was cataloged.

### **The Internet Archive and the Wayback Machine**

7. I am familiar with the Internet Archive, a nonprofit digital library established in 1996 that is accessible to the general public ([www.archive.org](http://www.archive.org)). Among its services, the Internet Archive maintains the “Wayback Machine,” an online archive that preserves and provides access to historical captures of webpages as they appeared on specific dates, thereby allowing researchers to track the development of websites over time. The Internet Archive has provided the Wayback Machine service since 2001. In my professional experience as a librarian, I have frequently used the Internet Archive’s Wayback Machine for research and instruction purposes, including teaching library patrons how to access archived webpages and relying on the Wayback Machine to locate otherwise hard-to-find online resources. I consider the Internet Archive’s recordskeeping practices, including its date-stamped captures and archiving protocols, to be rigorous and reliable, and comparable in their consistency and reliability to established bibliographic and cataloging standards such as MARC records, OCLC records and Dublin Core metadata.

## **The Computer Vision Foundation (CVF) and CVF open access**

8. I am familiar with Computer Vision Foundation (“CVF”), a nonprofit organization established to foster and support research in all aspects of computer vision. (<https://www.thecvf.com/>). CVF sponsors and co-sponsors leading international conferences in the field, including the IEEE Conference on Computer Vision and Pattern Recognition (“CVPR”) and the International Conference on Computer Vision (“ICCV”). As part of its mission, CVF maintains an “Open Access” website through which it makes the proceedings of these conferences publicly available without cost to researchers, practitioners, and members of the general public. (<https://openaccess.thecvf.com/menu>). According to CVF Open Access, “[e]xcept for the watermark, they [the research papers published] are identical to the accepted versions; the final published version of the proceedings is available on IEEE Xplore.” *Id.* CVF further explains that its objective is to ensure “timely dissemination of scholarly and technical work.” *Id.* In my professional experience, CVF’s open access repository is a widely used and authoritative source for conference papers and proceedings, and the papers posted there are freely accessible to anyone with an internet connection. CVF’s open access website allows users to locate conference papers by searching with a variety of parameters, including CVPR conference year, location, program chairs, workshop subject area, and title and author of each conference paper presented at corresponding conference.

## **IEEE Xplore**

9. I am also familiar with IEEE's Xplore digital library. In my experience, access to IEEE Xplore is widely provided by libraries. Many libraries, particularly university libraries, supplement their paper copies of articles and proceedings with digital IEEE Xplore subscriptions. Some libraries rely on IEEE Xplore records in lieu of their own catalog and MARC records. These online catalogs route users directly to the relevant IEEE Xplore page without displaying the library's catalog and MARC records to the user. IEEE's digital platform is also searchable using word search, key words, author names, conference names, and the like.

10. Not only do libraries rely on IEEE Xplore, but I also note that the U.S. Patent and Trademark Office uses IEEE Xplore in its prior art searches. *See* <https://www.uspto.gov/learning-and-resources/support-centers/scientific-and-technical-information-center-stic/electronic> (listing "Major Publishers of Electronic Non-Patent Literature Available at USPTO - Electronic Books," including 88,267 ebooks from IEEE Xplore; and listing "Major Publishers of Electronic Non-Patent Literature Available at USPTO - Electronic Journals" including 6,362 journals from IEEE Xplore) (last visited Sep. 16, 2025).

## **Ujiie**

11. I have reviewed Exhibit 1004 ("Ujiie"), which is a conference paper entitled "Approximated Prediction Strategy for Reducing Power Consumption of

Convolutional Neural Network Processor” by Takayuki Ujiie, et al. as published in the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.

12. I have reviewed the Declaration of Gordon MacPherson (Ex. 1010), the Director, Board Governance & IP Operations at The Institute of Electrical and Electronics Engineers, Inc. (“IEEE”). Mr. MacPherson testifies that he acts as a custodian of records for IEEE. Ex. 1010 ¶4. Attached to Mr. MacPherson’s declaration as Exhibit A is what he testifies is a true and correct copy of the Ujiie article available from IEEE Xplore that was maintained in the ordinary course of IEEE’s business. See Ex. 1010 ¶¶6-9. Based on that testimony, I refer to the IEEE provided copy of Ujiie as authenticated.

13. I note that Mr. MacPherson testifies that the IEEE’s copy of Ujiie was “published in 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), held on June 26, 2016 - July 1, 2016, in Las Vegas, NV, USA. Copies of the conference proceedings were made available no later than the last day of conference.” Ex. 1010 ¶13.

14. I compared IEEE’s copy of Ujiie to Exhibit 1004. Other than minor cosmetic differences, such as the header text on the first page and the footer formatting, the two are identical. Exhibit 1004 contains the same number of pages and the same substantive content as the authenticated version of Ujiie available through IEEE

Xplore. I note that Exhibit 1004's first page header states "[t]his CVPR workshop paper is the Open Access version, provide by the Computer Vision Foundation. Except for this watermark, it is identical to the version available on IEEE Xplore."

15. Thus, after comparing Exhibit 1004 with IEEE's true and correct copy of Ujiie, it is my opinion that each is an authentic version of Ujiie.

16. I located Ujiie myself from the Computer Vision Foundation (CVF) open access website by searching the year of the CVPR workshop, the subject matter Embedded Vision, title and author of Ujiie.

17. Attached hereto as Appendix UJIE01 is a screen capture of the Internet Archive Wayback Machine entry for [http://www.cvfoundation.org/openaccess/content\\_cvpr\\_2016\\_workshops/w14/papers/Ujiie\\_Approximated\\_Prediction\\_Strategy\\_CVPR\\_2016\\_paper.pdf](http://www.cvfoundation.org/openaccess/content_cvpr_2016_workshops/w14/papers/Ujiie_Approximated_Prediction_Strategy_CVPR_2016_paper.pdf). I secured these screen captures myself from [https://web.archive.org/web/20160805194848/http://www.cvfoundation.org/openaccess/content\\_cvpr\\_2016\\_workshops/w14/papers/Ujiie\\_Approximated\\_Prediction\\_Strategy\\_CVPR\\_2016\\_paper.pdf](https://web.archive.org/web/20160805194848/http://www.cvfoundation.org/openaccess/content_cvpr_2016_workshops/w14/papers/Ujiie_Approximated_Prediction_Strategy_CVPR_2016_paper.pdf). Accordingly, Appendix UJIE01 is a screen capture of the Internet Wayback Machine showing Ujiie (Ex. 1004) as published on the CVF open access website, with capture date of August 5, 2016.

18. Attached hereto as Appendix UJIE02 is a screen capture of the Internet Archive Wayback Machine entry for [http://www.cvfoundation.org/openaccess/content\\_cvpr\\_2016\\_workshops/w14/papers/Ujiie\\_Approximated\\_Prediction\\_Strategy\\_CVPR\\_2016\\_paper.pdf](http://www.cvfoundation.org/openaccess/content_cvpr_2016_workshops/w14/papers/Ujiie_Approximated_Prediction_Strategy_CVPR_2016_paper.pdf).

foundation.org/openaccess/CVPR2016\_workshops/menu.py. I secured these screen captures myself from [https://web.archive.org/web/20160909133320/http://www.cv-foundation.org/openaccess/CVPR2016\\_workshops/menu.py](https://web.archive.org/web/20160909133320/http://www.cv-foundation.org/openaccess/CVPR2016_workshops/menu.py). This record was captured as of September 9, 2016.

19. Appendix UJIIE02 is an Internet Archive capture of a web page entitled Computer Vision Foundation Open Access as hosted on the CV Foundation website. This page describes the various conference proceedings associated with the 2016 Conference on Computer Vision and Pattern Recognition, each associated with a workshop held as part of the conference. The PDF copy of the Ujiie paper presented in Exhibit 1004 was secured by clicking the 'Embedded Vision' link present on page 2 of this appendix and reviewing the results.

20. Attached hereto as Appendix UJIIE03 is a screen capture of the Internet Archive Wayback Machine entry for <http://cvisioncentral.com/promotion/evw2016/>. I secured these screen captures myself from <https://web.archive.org/web/20160924044115/http://cvisioncentral.com/promotion/evw2016/>. This record was captured as of September 24, 2016.

21. Appendix UJIIE03 is an Internet Archive capture of a web page entitled The Twelfth IEEE Embedded Vision Workshop including the Program as hosted on the Computer Vision Central website. This page presents a list of papers presented at

the Embedded Vision workshop at the 2016 Conference on Computer Vision and Pattern Recognition. According to an entry found on page 2 of Appendix UJIIE05, the Ujiie paper was presented at this workshop as of July 1, 2016.

22. Attached hereto as Appendix UJIIE04 is a screen capture of the Internet Archive Wayback Machine entry for [http://cvpr2016.thecvf.com/program/industry\\_expo](http://cvpr2016.thecvf.com/program/industry_expo). I secured these screen captures myself from [https://web.archive.org/web/20160817110953/http://cvpr2016.thecvf.com/program/industry\\_expo](https://web.archive.org/web/20160817110953/http://cvpr2016.thecvf.com/program/industry_expo). This record was captured as of August 17, 2016.

23. Appendix UJIIE04 is an Internet Archive capture of a web page entitled CVPR 2016 Industry Expo as hosted on the CVPR2016.thecvf.com website. This page describes the attendance of the 2016 Conference on Computer Vision and Pattern Recognition with the following passage: ‘Over 3,000 CVPR attendees will create a one-of-a-kind opportunity for networking, recruiting, inspiration and motivation’. This page also indicates over 50 companies exhibited at this conference, including industry leaders such as Adobe, Amazon, Apple, Elsevier, Google, Intel, IBM, Microsoft, NVIDIA, Sony and Springer.

24. Based on the foregoing accessibility, searchability, and cataloging provided by the CVF’s open access during the time frame in which Ujiie was available on that platform, it is my determination that a person interested in and ordinarily skilled in

the subject matter of Ujiie, or a researcher or librarian interested in the same, could locate it through the exercise of ordinary diligence as of at least August 05, 2016. For example, a simple search of the CVPR 2016 open access website using keywords from the title, such as “Approximated Prediction Strategy for Reducing Power Consumption” and/or “Convolutional Neural Network Processor,” retrieves the Ujiie paper.

25. Attached hereto as Appendix UJIE05 is a true and correct copy of the MARC record for a serial entry of “IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops” as held by the Penn State University Library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix UJIE01 accurately describes the title, publisher and ISSN number of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. *Id.*, 2. The “Showing Results” field present on page 3 of this record indicates the library’s collection includes workshops from 1994 until present day, which necessarily includes the 2016 proceedings containing the Ujiie paper.

26. The 008 field of the MARC record in Appendix UJIE01 indicates the date of record creation. The 008 field of Appendix UJIE01 indicates the Penn State University Library first acquired IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops as of March 31, 2011 and has continued

to hold the series through the present. The 310 field of the MARC record indicates the current publication frequency. The 310 field of Appendix UJIIE01 indicates that IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops is updated annually, consistent with CVPR's yearly conference schedule. The 035/OCLC number indicates the unique OCLC identified, confirming this entry is in WorldCat and traceable across other libraries. The 035 field of this record shows multiple control numbers assigned by different bibliographic systems, including OCLC (WorldCat) and Serials Solutions (WaSeSS). These identifiers confirm that the record for the CVPR Workshops proceedings was not only cataloged at Penn State University but also shared across widely used bibliographic databases, making it discoverable to researchers worldwide. Considering this information, it is my determination that the 2016 proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, including Ujiie, was made available to the public shortly after its initial publication in 2016.

### **Moons**

27. I have reviewed Exhibit 1005 ("Moons"), "Energy-Efficient ConvNets Through Approximate Computing" by Bert Moons, et al. as published in Proceedings of the 2016 Winter Conference on Applications of Computer Vision.

28. I have reviewed the Declaration of Gordon MacPherson (Ex. 1010), the Director, Board Governance & IP Operations at The Institute of Electrical and Electronics Engineers, Inc. (“IEEE”). Mr. MacPherson testifies that he acts as a custodian of records for IEEE. Ex. 1010 ¶4. Attached to Mr. MacPherson’s declaration as Exhibit B is what he testifies is a true and correct copy of the Moons article available from IEEE Xplore that was maintained in the ordinary course of IEEE’s business. See Ex. 1010 ¶¶6-8, 10. Based on that testimony, I refer to the IEEE provided copy of Moons as authenticated.

29. I note that Mr. MacPherson testifies that the IEEE’s copy of Moon was “published in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), held on March 7-10, 2016, in Lake Placid, NY, USA. Copies of the conference proceedings were made available no later than the last day of conference. This article was added to the IEEE digital library Xplore on May 26, 2016, as shown from the abstract from IEEE Xplore, and was publicly available for download from IEEE Xplore by May 26, 2016, and is currently available for public download from the IEEE digital library, IEEE Xplore.” Ex. 1010 ¶14.

30. I compared IEEE’s copy of Moons to Exhibit 1005. Other than minor cosmetic differences the two are identical. Exhibit 1005 contains the same number of pages and the same substantive content as the authenticated version of Moons available through IEEE Xplore.

31. Thus, after comparing Exhibit 1005 with IEEE's true and correct copy of Moons, it is my opinion that each is an authentic version of Moons.

32. Attached hereto as Appendix MOONS01 is a true and correct copy of the MARC record for Proceedings of the 2016 Winter Conference on Applications of Computer Vision as held by the Penn State University Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix MOONS01 accurately describes the title, publisher and ISBN number of Proceedings of the 2016 Winter Conference on Applications of Computer Vision.

33. Attached hereto as Appendix MOONS02 is a true and correct copy of "Energy-Efficient ConvNets Through Approximate Computing" as published in Proceedings of the 2016 Winter Conference on Applications of Computer Vision as held by the Penn State University Library. Appendix MOONS02 was secured from the library's collection. In comparing Exhibit 1005 to Appendix MOONS02, it is my determination that Exhibit 1005 is a true and correct copy of "Energy-Efficient ConvNets Through Approximate Computing" by Bert Moons, et al as published in Proceedings of the 2016 Winter Conference on Applications of Computer Vision.

34. The 008 field of the MARC record in Appendix MOONS01 indicates the date of record creation. The 008 field of Appendix MOONS01 indicates the Penn State University Library first acquired Proceedings of the 2016 Winter Conference on Applications of Computer Vision as of February 17, 2016. Considering this

information, it is my determination that Proceedings of the 2016 Winter Conference on Applications of Computer Vision and therefore “Energy-Efficient ConvNets Through Approximate Computing” was made available to the public shortly after February 17, 2016.

### **Kaul**

35. I have reviewed Exhibit 1006 (“Kaul”), “A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS” by Himanshu Kaul, et al as published in Proceedings of the 2016 ISSCC.

36. I have reviewed the Declaration of Gordon MacPherson (Ex. 1010), the Director, Board Governance & IP Operations at The Institute of Electrical and Electronics Engineers, Inc. (“IEEE”). Mr. MacPherson testifies that he acts as a custodian of records for IEEE. Ex. 1010 ¶4. Attached to Mr. MacPherson’s declaration as Exhibit C is what he testifies is a true and correct copy of the Moons article available from IEEE Xplore that was maintained in the ordinary course of IEEE’s business. See Ex. 1010 ¶¶6-8, 11. Based on that testimony, I refer to the IEEE provided copy of Moons as authenticated.

37. I note that Mr. MacPherson testifies that the IEEE’s copy of Kaul was “published in 2016 IEEE International Solid-State Circuits Conference (ISSCC), held on January 31, 2016 - February 4, 2016, in San Francisco, CA, USA. Copies of

the conference proceedings were made available no later than the last day of conference. This article was added to the IEEE digital library Xplore on February 25, 2016, as shown from the abstract from IEEE Xplore, and was publicly available for download from IEEE Xplore by February 25, 2016, and is currently available for public download from the IEEE digital library, IEEE Xplore.” Ex. 1010 ¶15.

38. I compared IEEE’s copy of Kaul to Exhibit 1006. Other than minor cosmetic differences the two are identical. Exhibit 1006 contains the same number of pages and the same substantive content as the authenticated version of Kaul available through IEEE Xplore.

39. Thus, after comparing Exhibit 1006 with IEEE’s true and correct copy of Kaul, it is my opinion that each is an authentic version of Kaul.

40. Attached hereto as Appendix KAUL01 is a true and correct copy of the MARC record for Proceedings of the 2016 ISSCC as held by the Penn State University Library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix KAUL01 accurately describes the title, publisher and ISBN number of Proceedings of the 2016 ISSCC.

41. Attached hereto as Appendix KAUL02 is a true and correct copy of “A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS” as published in Proceedings of the 2016 ISSCC as held by the Penn State University Library.

Appendix KAUL02 was secured from the library's collection. In comparing Exhibit 1006 to Appendix KAUL02, it is my determination that Exhibit 1006 is a true and correct copy of "A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS" by Himanshu Kaul, et al. as published in Proceedings of the 2016 ISSCC.

42. The 008 field of the MARC record in Appendix KAUL01 indicates the date of record creation. The 008 field of Appendix KAUL01 indicates the Penn State University Library first acquired Proceedings of the 2016 ISSCC as of February 17, 2016. Considering this information, it is my determination that Proceedings of the 2016 ISSCC and therefore "A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS" was made available to the public shortly after February 17, 2016.

### **Compensation and Declaration**

43. I have been retained on behalf of the Petitioner to provide assistance in the above-illustrated matter in establishing the authenticity and public availability of the documents discussed in this declaration. I am being compensated for my services in this matter at the rate of \$200.00 per hour plus reasonable expenses. My statements are objective, and my compensation does not depend on the outcome of this matter.

44. I declare under penalty of perjury that the foregoing is true and correct. I hereby declare that all statements made herein of my own knowledge are true and

that all statements made on information and belief are believed to be true; and further that these statements were made the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: 09/23/2025

A handwritten signature in cursive script, appearing to read "June Ann Munford".

June Ann Munford

# APPENDIX CV

June A. Munford  
Curriculum Vitae

## **Education**

University of Wisconsin-Milwaukee - MS, Library & Information Science, 2009  
Milwaukee, WI

- Coursework included cataloging, metadata, data analysis, library systems, management strategies and collection development.
- Specialized in library advocacy, cataloging and public administration.

Grand Valley State University - BA, English Language & Literature, 2008  
Allendale, MI

- Coursework included linguistics, documentation and literary analysis.
- Minor in political science with a focus in local-level economics and government.

## **Professional Experience**

Researcher / Expert Witness, October 2017 – present  
Freelance ● Pittsburgh, Pennsylvania & Grand Rapids, Michigan

- Material authentication and public accessibility determination. Declarations of authenticity and/or public accessibility provided upon research completion. Experienced with appeals and deposition process.
- Research provided on topics of public library operations, material publication history, digital database services and legacy web resources.
- Past clients include Alston & Bird, Arnold & Porter, Baker Botts, Fish & Richardson, Erise IP, Irell & Manella, O'Melveny & Myers, Perkins-Coie, Pillsbury Winthrop Shaw Pittman and Slayden Grubert Beard.

Library Director, February 2013 - March 2015  
Dowagiac District Library ● Dowagiac, Michigan

- Executive administrator of the Dowagiac District Library. Located in Southwest Michigan, this library has a service area of 13,000, an annual operating budget of over \$400,000 and total assets of approximately \$1,300,000.

- Developed careful budgeting guidelines to produce a 15% surplus during the 2013-2014 & 2014-2015 fiscal years while being audited.
- Using this budget surplus, oversaw significant library investments including the purchase of property for a future building site, demolition of existing buildings and building renovation projects on the current facility.
- Led the organization and digitization of the library's archival records.
- Served as the public representative for the library, developing business relationships with local school, museum and tribal government entities.
- Developed an objective-based analysis system for measuring library services - including a full collection analysis of the library's 50,000+ circulating items and their records.

November 2010 - January 2013

Librarian & Branch Manager, Anchorage Public Library ● Anchorage, Alaska

- Headed the 2013 Anchorage Reads community reading campaign including event planning, staging public performances and creating marketing materials for mass distribution.
- Co-led the social media department of the library's marketing team, drafting social media guidelines, creating original content and instituting long-term planning via content calendars.
- Developed business relationships with The Boys & Girls Club, Anchorage School District and the US Army to establish summer reading programs for children.

June 2004 - September 2005, September 2006 - October 2013

Library Assistant, Hart Area Public Library

Hart, MI

- Responsible for verifying imported MARC records and original MARC cataloging for the local-level collection as well as the Michigan Electronic Library.
- Handled OCLC Worldcat interlibrary loan requests & fulfillment via ongoing communication with lending libraries.

## **Professional Involvement**

Alaska Library Association - Anchorage Chapter

- Treasurer, 2012

#### Library Of Michigan

- Level VII Certification, 2008
- Level II Certification, 2013

#### Michigan Library Association Annual Conference 2014

- New Directors Conference Panel Member

#### Southwest Michigan Library Cooperative

- Represented the Dowagiac District Library, 2013-2015

### **Professional Development**

#### Library Of Michigan Beginning Workshop, May 2008

##### Petoskey, MI

- Received training in cataloging, local history, collection management, children's literacy and reference service.

#### Public Library Association Intensive Library Management Training, October 2011

##### Nashville, TN

- Attended a five-day workshop focused on strategic planning, staff management, statistical analysis, collections and cataloging theory.

#### Alaska Library Association Annual Conference 2012 - Fairbanks, February 2012

##### Fairbanks, AK

- Attended seminars on EBSCO advanced search methods, budgeting, cataloging, database usage and marketing.

### **Depositions**

#### 2019 ● Fish & Richardson

Apple v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282 and 39521-00421IP2)

#### 2019 ● Erise IP

Implicit, LLC v. Netscout Systems, Inc (Case No. 2:18-cv-53-JRG)

#### 2019 ● Perkins-Coie

Adobe Inc. v. RAH Color Technologies LLC (Case No. IPR2019-00627, IPR2019-00628, IPR2019-00629 and IPR2019-00646)

2020 ● O'Melveny & Myers

Maxell, Ltd. v. Apple Inc. (Case No. 5:19-cv-00036-RWS)

2021 ● Pillsbury Winthrop Shaw Pittman LLP

Intel v. SRC (Case No. IPR2020-1449)

2022 ● Perkins-Coie

Realtek v. Future Link (Case No. IPR2021-01182)

2023 ● Fish & Richardson

Neuroderm Ltd. v. Abbvie, Inc (Case No. PGR2022-00040)

2023 ● Fish & Richardson

Nearmap US Inc. v. Pictometry International Corp. (Case No. IPR2022-00735)

2023 ● Fish & Richardson

Samsung Electronics v. MemoryWeb LLC (Case No. 39843-0136PS1)

2023 ● Pillsbury Winthrop Shaw Whitman LLP

Gravel Rating Systems v. Costco Wholesale Corp. (Case No. 4:21-cv-149-ALM)

2024 ● Willkie-Farr

Netflix, Inc. v. VideoLabs. Inc. (Case No. IPR2023-00628)

2024 ● Quinn Emanuel Urquhart & Sullivan

Purdue University v. Wolfsped (Case No. 1:21-CV-840)

2024 ● Fish & Richardson

Dish Network v. Entropic Communications (Case No. IPR2024-00393)

2025 ● Perkins Coie

Amazon.com, Inc. v. Nokia Technologies Oy (Case No. IPR2024-00572)

2025 ● Fish & Richardson

Dish Network v. Entropic Communications (Case No. IPR2024-00373)

### **Limited Case and Clientele History**

Alston & Bird

- Ericsson

v. Collision Communications (Case No. IPR2022-01233)

- Nokia

v. Neptune Subsea, Xtera (Case No. 1:17-cv-01876)

- Universal Electronics Inc

v. Roku Inc (Case No. IPR2022-00818)

Alavi & Anaipakos PLLC

- Stingray Group Inc. and Stingray Music USA, Inc.

v. Edwin A. Hernandez-Mondragon and Eglia Corp. (Case No. 1:24-cv-21226-RAR)

Arnold & Porter

- Ivantis

v. Glaukos (Case No. 8:18-cv-00620)

- Samsung

v. Jawbone (Case No. 2:21-cv-00186)

Benesch Friedlander Coplan & Aronoff

- Voyis  
v. Cathx (Case No. 5:21-cv-00077-RWS)

Buchanan, Ingersoll & Rooney PC

- Google LLC, AT&T Service Inc., T Mobile USA Inc., Cellco Partnership, Ericsson Inc. & Nokia of America Corp.  
v. KT Corp & Pegasus Wireless Innovation LLC (Case No. IPR2025-00293)

Deschert LLP

- Smaxtec Animal Care  
v. ST Reproductive Technologies, LLC (Case No. IPR2024-00885)

Duane Morris

- Kangxi Communication Technologies  
v. Skyworks Solutions Inc. (Case No. IPR2025-00373)
- Microsoft  
v. Edge Networking Systems LLC (Case No. 1:24-cv-00215)

Erise I.P.

- Apple  
v. Ericsson Inc. (Case No. IPR2022-00715)  
v. Future Link Systems (Case No. IPRs 6317804, 6622108, 6807505, and 7917680)  
v. INVT (Case No. 20-1881)  
v. Navblazer LLC (Case No. IPR2020-01253)  
v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)

- v. Quest Nettech Corp (Case No. 2:19-cv-00118-JRG)
- v. Telefonaktiebolaget LM Ericsson (Case No. IPR2022-00275)
- v. Theta IP, LLC (Case No. IPR2024-00818)
- v. THL Holding Company LLC (Case No. 1:23-cv-00548)

- Fanduel

- v. CGT (Case No. 19-1393)

- Garmin

- v. Phillips North America LLC (Case No. 2:19-cv-6301-AB-KS)

- Netscout

- v. Longhorn HD LLC (Case No. 2:20-cv-00349)
  - v. Implicit, LLC (Case No. 2:18-cv-53-JRG)

- Sony Interactive Entertainment LLC

- v. Bot M8 LLC (Case No. IPR2020-01288)
  - v. Infernal Technology LLC (Case No. 2:19-CV-00248-JRG)

- Tesla

- v. Charge Fusion Technologies LLC (Case No. IPR2025-00032)

- Unified Patents

- v. GE Video Compression (Case No. 2:19-cv-248)

## Fish & Richardson

- Apple

- v. AliveCor (Case No. 3:21-cv-03958)
  - v. LBS Innovations (Case No. 2:19-cv-00119-JRG-RSP)
  - v. Koss Corporation (Case No. IPR2021-00305)

v. Masimo (Case No. IPR 50095-0012IP1, 50095-0012IP2, 50095-0013IP1, 50095-0013IP2, 50095-0006IP1, 50095-0135IP1)

v. Neonode (Case No. 21-cv-08872-EMC)

v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)

v. Resonant Sys. (Case No. 7:23-cv-00077-ADA)

- AsusTek Computer Inc.

v. Videolabs, Inc. (Case No. 22-CV-00720-ADA)

- Dell

v. Neo Wireless (Case No. IPR2022-00616)

- Dish Network

v. Entropic Communications, LLC (Case No. 2:2023-CV-01043)

v. Entropic Communications LLC (Case No. IPR2024-00393)

v. Realtime Adaptive Streaming (Case No. 1:17-CV-02097-RBJ)

v. TQ Delta LLC (Case No. 18-1798)

- Evapco Dry Cooling

v. SPG Dry Cooling (Case No. IPR2021-00688)

- Genetec

v. Sensormatic Electronics (Case No. 1:20-CV-00760)

- Huawei

v. Bell Northern Research LLC (Case No. IPR2019-01174)

- Kianxis

v. Blue Yonder (Case No. 3:20-cv-03636)

- LG Electronics
  - v. Bell Northern Research LLC (Case No. 3:18-cv-2864-CAB-BLM)
- MediaTek
  - v. MOSAID (Case No. IPR2024-00718)
- Metaswitch
  - v. Sonus Networks (Case No. IPR2018-01719)
- Microsoft
  - v. Throughputer Inc (Case No. IPR2022-00757)
- Mom Enterprises
  - v. Ddrops Company (Case No. 1:22-cv-00332)
- MLC Intellectual Property
  - v. MicronTech (Case No. 3:14-cv-03657-SI)
- Nearmap Inc
  - v. EagleView Technologies (Case No. IPR2022-01009)
- Neuroderm Ltd.
  - v. Abbvie, Inc (Case No. PGR2022-00040)
- Posco Co Ltd.
  - v. Pascal Drillet et al. (Case No. IPR2025-00370)
- Realtek Semiconductor
  - v. Future Link (Case No. IPR2021-01182)

- Quectel
  - v. Koninklijke Philips (Case No. 1:20-cv-01710)
- Samsung
  - v. Aire Technology (Case No. IPR2022-00877)
  - v. Bell Northern Research (Case No. 2:19-cv-00286-JRG)
  - v. Communication Technologies Inc (Case No. IPR2022-01221)
  - v. Jawbone Innovations (Case No. IPR2022-00865)
  - v. MemoryWeb LLC (Case No. IPR2022-00885)
  - v. Dodots Licensing Solutions (Case No. IPR2023-00701)
  - v. Telefonaktiebolaget LM Ericsson (Case No. IPR2021-00615)
- Texas Instruments
  - v. Vantage Micro LLC (Case No. IPR2020-01261)
- Xilinx
  - v. Sentient Sensors LCC (Case No. 1:22-cv-00173)

Goodwin Proctor

- Intel
  - v. Proxense, LLC (Case No. 6:24-cv-00283)
- Samsung Electronics
  - v. Mobile Data Technologies (Case No. 2:24-cv-00435)
- Taiwan Semiconductor Manufacturing Company Ltd.
  - v. Advanced Integrated Circuit Process LLC (Case No. 2:24-cv-00623)

Groombridge Wu

- Nearmap

v. Eagle View Technologies Inc. and Pictometry International Corp. (Case No. 2:21-cv-00283-TS-DAO)

Irell & Manella

- Curium

Latham & Watkins LLP

- Apple

v. Advanced Coding Research LLC

Leydig, Voit & Mayer, Ltd.

- Public Availability Consultancy

O'Melveny & Myers

- Apple

v. Maxell (Case No. 5:19-cv-00036-RWS)

- Disney Media & Entertainment Distribution LLC

v. Digital Media Technology Holdings, LLC (Case No. 2:22-cv-01642)

- Micron Technology Inc.

v. Besang Inc (Case No. IPR2023-00900)

- Samsung

v. Daedalus Prime LLC (1335 Investigation)

Perkins-Coie

- Amazon
  - v. Nokia (Case No. IPR2024-00847 and IPR2024-00848)
  - v. Nokia Technologies Oy (Case No. IPR2024-00572)
- Heru Industries
  - v. The UAB Research Foundation (Case No. IPR2022-01148)
- Intel Corporation
  - v. BeSang Inc. (Case No. IPR2023-00991)
  - v. BitMICRO (Case No. 5:23-cv-00625)
  - v. Telefonaktiebolaget L M Ericsson (Case No. IPR2024-00610)
- Realtek Semiconductor
  - v. Future Link (Case No. IPR2021-01182)
- r-pac
  - v. Adasa, Inc. (Case No. 1:2024cv06102)
- Twitter Inc
  - v. VOIP-Pal.com (Case No. 3:20-cv-02397-JD)
- TCL Industries
  - v. Koninklijke Philips NV (Case No. IPR2021-00495, IPR2021-00496 and IPR2021-00497)
- VusionGroup
  - v. Hanshow (Case No. IPR2024-00857)

Pillsbury Winthrop Shaw Pittman

- Intel

v. FG SRC LLC (Case No. 6:20-cv-00315)

- Gravel Rating Systems

v. Costco (Case No. 4:21-cv-149)

v. Lowe's Home Centers (Case No. 4:21-cv-150)

v. T-Mobile USA (Case No. 4:21-cv-152)

v. Kohl's Inc. (Case No. 4:21-cv-258)

v. Under Armor (Case No. 4:21-cv-356)

Quinn Emanuel

- Exact Sciences Corporation

v. Geneoscopy, Inc. (Case No. IPR2024-00459 and IPR2024-1330)

- ServiceNow, Inc.

v. InQuisient, Inc. (Case No. 22-900-CJB)

- Wolfsped

v. The Trustees of Purdue University (Case No. 1:2021-cv-00840)

Sheppard, Mullin, Richter & Hampton LLP

- Advanced Micro Devices, Inc and Pensando Systems, Inc.

v. Concurrent Ventures LLC and Xtreamedge, Inc. (Case No. IPR2025-00478)

- Cadence Design Systems Inc.

v. Semiconductor Design Technologies (Case No. 3:23-cv-01001)

Shook, Hardy & Bacon LLP

- HP Inc.

v. Universal Connectivity Technologies (Case No. 1:23-cv-01177)

Stern, Kessler, Goldstein & Fox

- Preliminary Research

Wilmer, Cutler, Pickering, Hale and Dorr

- Apple

v. Shunock (Case No. 1:23-cv-08598)

- Quest Diagnostics Inc., Haystack Oncology Inc. and Haystack Oncology GmbH

v. Natera, Inc. (Case No. 1:23-cv-08598)

- Roche Diabetes, Inc.

v. Trividia Health, Inc. (Case No. 1:24-cv-00668)

Willkie, Farr & Gallagher LLP

- Lenovo, Dell, HP

v. Universal Connectivity Technologies Inc. (Case No. 2:2023cv00449)

- Neurent

v. The Foundry, LLC (Case No. IPR2024-00669)

- Netflix, Inc.

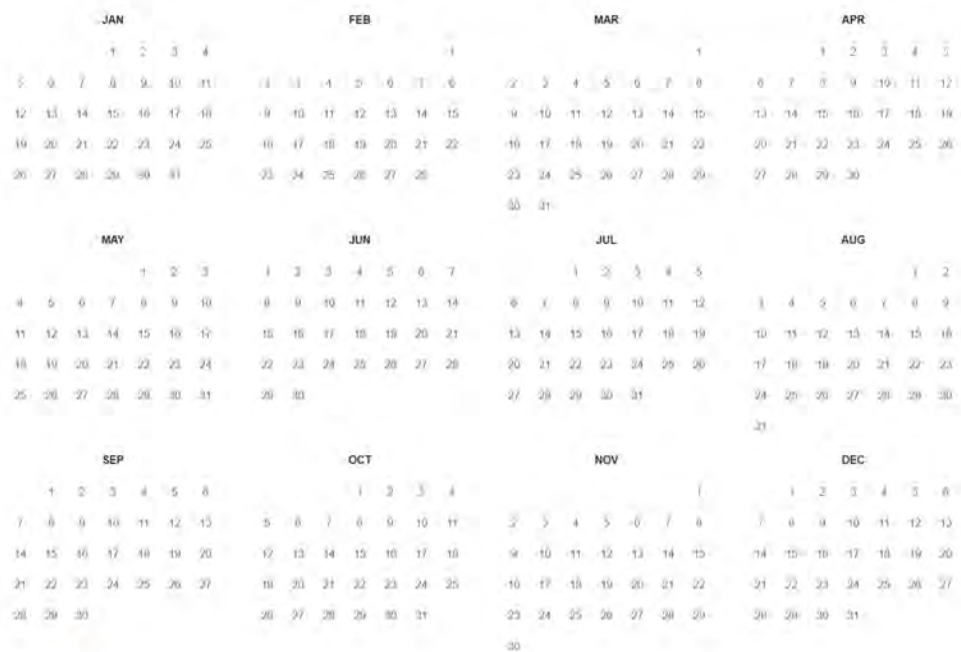
v. VideoLabs, Inc. (Case No. IPR2023-00628)

# APPENDIX UJIIIE01

INTERNET ARCHIVE  
WayBackMachine  
Explore more than 946 billion web pages saved over time  
http://www.cv-foundation.org/openaccess/content\_cvpr\_2016\_w

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved 6 times between August 5, 2016 and April 16, 2024.



### Note

This calendar view maps the number of times [http://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016\\_workshops/w14/papers/Ujii Approximated Prediction Strategy CVPR\\_2016\\_paper.pdf](http://www.cv-foundation.org/openaccess/content_cvpr_2016_workshops/w14/papers/Ujii Approximated Prediction Strategy CVPR_2016_paper.pdf) was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library and Archive-It.org.  
Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor

1 / 7 100% +

**CVF** This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the version available on IEEE Xplore.

## Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor

Takayuki Ujiie Masayuki Hiromoto Takashi Sato  
Graduate School of Informatics, Kyoto University  
Yoshida-hon-machi, Sakyo, Kyoto, Japan  
paper@easter.kuee.kyoto-u.ac.jp

### Abstract

*Convolutional neural network (CNN) is becoming popular because of its great ability for accurate image recognition. However, the computational cost is extremely high, which increases power consumption of embedded CV systems. This paper proposes an efficient computing method, LazyConvPool (LCP), and its hardware architecture to reduce power consumption of CNN-based image recognition.*

work (CNN) is one of those representation learning methods and has achieved remarkable results in various image recognition tasks [10, 15, 17]. CNN computes 2D convolutions on images over and over again to extract the features. The local features of images are acquired by adjusting the weights of the convolution filters. This general framework makes it possible for CNN to accurately recognize various classes of images without giving the manually designed features. In order to achieve even higher accuracy, the network

# Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor

Takayuki Ujiie    Masayuki Hiromoto    Takashi Sato  
Graduate School of Informatics, Kyoto University  
Yoshida-hon-machi, Sakyo, Kyoto, Japan  
paper@easter.kuee.kyoto-u.ac.jp

## Abstract

*Convolutional neural network (CNN) is becoming popular because of its great ability for accurate image recognition. However, the computational cost is extremely high, which increases power consumption of embedded CV systems. This paper proposes an efficient computing method, LazyConvPool (LCP), and its hardware architecture to reduce power consumption of CNN-based image recognition. The LCP exploits redundancy of operations in CNN and only executes essential convolutions by an approximated prediction technique. We also propose Sign Connect, which is a low computational-cost approximated prediction without any multiplications. The experimental evaluation using image classification dataset shows that the proposed method reduces the power consumption by 17.8%–20.2% and energy consumption by 11.4%–14.1% while retaining recognition performance.*

## 1. Introduction

Image recognition is a key technology in various fields such as robotics, automotive, security, and medical applications. Realizing efficient embedded systems for such applications is nontrivial because we must co-design both algorithm and hardware architecture. A large number of dedicated vision processors have been proposed, which utilize handcrafted features (Haar-like features [22], SIFT [20], HOG [8], etc.) and learning algorithms (SVM [6], Boosting [9], etc.). However, the design of the specialized processor is costly in terms of development efforts albeit it could potentially achieve good performance.

In contrast to the traditional recognition algorithms that utilize the handcrafted features, representation learning [1] is becoming increasingly popular in the state-of-the-art recognition methods. The key idea of the representation learning is to let the system learn the feature by itself, in addition to learning the classifier. Convolutional neural net-

work (CNN) is one of those representation learning methods and has achieved remarkable results in various image recognition tasks [10, 15, 17]. CNN computes 2D convolutions on images over and over again to extract the features. The local features of images are acquired by adjusting the weights of the convolution filters. This general framework makes it possible for CNN to accurately recognize various classes of images without giving the manually designed features. In order to achieve even higher accuracy, the network structure of the CNN tends to be made deeper and deeper, requiring huge computational resources. An efficient computing scheme to facilitate such a large number of convolutions is desired to realize accurate image recognition on energy-limited embedded systems.

Among the several kinds of approaches for realizing energy-efficient CNN, recent researches focus on the inherent resilience of CNN [21, 23]. Calculation of pattern recognition algorithms generally contains redundant calculation to ensure the robustness against noisy inputs. In other words, the redundancy in most of the recognition algorithms can be pursued to reduce computational resources. In this work, we focus on the characteristic of the max pooling operation, which is an integral component of CNN that selects a representative feature in a small region. Specifically, we propose LazyConvPool (LCP), a new computation method for CNN's convolution layers and pooling layers. In LCP, only the essential convolutions that affect the final result are executed on the basis of the prediction results by approximated computations. The key features of the proposed LCP are as follows:

- LCP predicts a local region that will be selected later in the pooling layer to limit the application of the accurate and thus expensive convolutional operations to the predicted region.
- The prediction is based on the newly proposed *Sign Connect*, which gives a significantly lightweight approximation of the convolution with no multiplication operation, while keeping prediction accuracy.

- An efficient hardware architecture is also proposed to implement a low power LCP-based CNN with a minimal overhead.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the proposed methods, LCP and Sign Connect. Section 4 presents the hardware architecture for the proposed method and shows an implementation details to improve performance. Section 5 shows the experimental results, and finally, Section 6 concludes the paper.

## 2. Related works

### 2.1. Convolutional neural network

A typical CNN consists of convolution layers, pooling layers, and a fully connected layer as shown in Figure 1. First, in the convolution layers, 2D convolution is computed as

$$W * x(m, n) = \sum_u \sum_v W(u, v) x(m + u, n + v), \quad (1)$$

where  $x$  is an input image and  $W$  is a weight matrix of the convolution filter. The operator ‘\*’ means 2D convolution. This layer extracts local features of images to generate feature maps by using a large number of different filters. Then in the pooling layers, a representative feature value is selected from a small region called a pooling window. The pooling layer is useful for translation invariance and reduction of computational costs. As the representative value, maximum (max) or average values are commonly used. In this paper, we use max pooling, which selects the max value from the pooling window. After the convolution and the pooling, non-linearity is used as an activation function similarly to the traditional neural networks. Here, piecewise continuous functions are generally used as the activation function. In practice, rectified linear unit (ReLU) is widely used because it does not saturate gradients while requiring a small amount of computation. ReLU is represented as

$$f(x) = \max(0, x). \quad (2)$$

In the fully connected layer, an output  $\mathbf{x}^{(m)}$  is computed from the outputs of the previous layer  $\mathbf{x}^{(m-1)}$  as

$$\mathbf{x}^{(m)} = f(W^T \mathbf{x}^{(m-1)} + \mathbf{b}), \quad (3)$$

where  $W$  is a weight matrix,  $\mathbf{b}$  is a bias vector in this layer, and  $f(\cdot)$  is an activation function. At the final stage of the network, the softmax function is often used as the activation function for multi-class classification:

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (i = 0, \dots, n-1). \quad (4)$$

The softmax function has an advantage because it normalizes values in the final stage and facilitates error computation in the training phase. The number of the output units in the final stage is equal to the number of the labels, and the classification result  $y$  is given by

$$y = \arg \max_i x_i. \quad (5)$$

### 2.2. Computationally efficient neural networks

CNN requires large computational costs as well as other neural networks. It is because a huge number of multiplications are executed to propagate signals in the networks. Some recent works are trying to tackle this problem by replacing the multiplications with simpler operations. Network pruning [14] and network compression [2, 3] reduce the amount of weight parameters by compressing networks. These methods try to compress networks by utilizing some representative values to describe the whole weight matrix. On the other hand, BinaryNet [7] and Ternary Connect [19] are the methods that limit the weight values to  $\{-1, +1\}$  or  $\{-1, 0, +1\}$  during the training phase in order to reduce the memory usage and computation complexity. In the following paragraph, we describe the more details of Ternary Connect, on which our proposed method Sign Connect is based.

**Ternary Connect** Ternary Connect is a computation method for neural networks that eliminates multiplications in the propagation. Ternary Connect replaces the original real weight  $w(u, v) \in [-1, 1]$  by an alternative weights  $W(u, v) \in \{-1, 0, 1\}$ , which is stochastically determined by  $w(u, v)$ . If  $w(u, v) > 0$ , the probability of the  $W(u, v)$  value is given by

$$\begin{cases} P(W(u, v) = 1) &= w(u, v) \\ P(W(u, v) = 0) &= 1 - w(u, v) \end{cases} \quad (6)$$

Otherwise, if  $w(u, v) < 0$ , the probability is given by

$$\begin{cases} P(W(u, v) = -1) &= -w(u, v) \\ P(W(u, v) = 0) &= 1 + w(u, v) \end{cases} \quad (7)$$

By using  $W(u, v)$  in propagation computation in stead of  $w(u, v)$ , multiplications are completely omitted.

Although Ternary Connect is very useful to reduce computation in the training phase, it can deteriorate recognition performance if it is simply applied to the classification phase. In addition, a suitable hardware architecture should be considered at the same time in order to receive the benefit of reducing multiplications.

### 2.3. CNN processors

CNN in embedded systems has been actively researched in several aspects. As application-specific processors, some

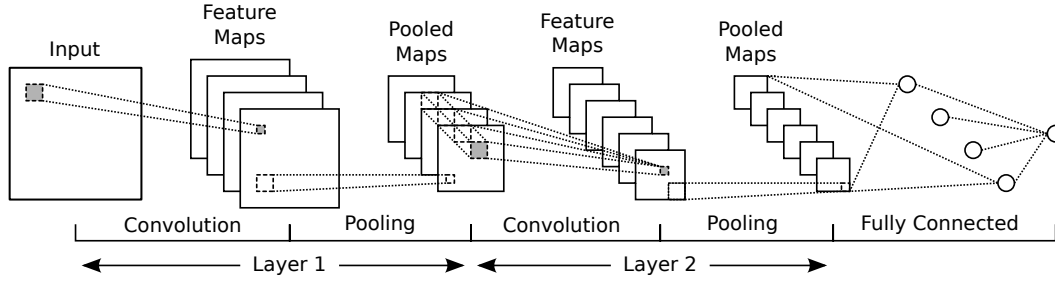


Figure 1: A typical architecture of a convolutional neural network. It consists of multiple sets of convolution and pooling layers and succeeding fully connected multi-layer perceptrons.

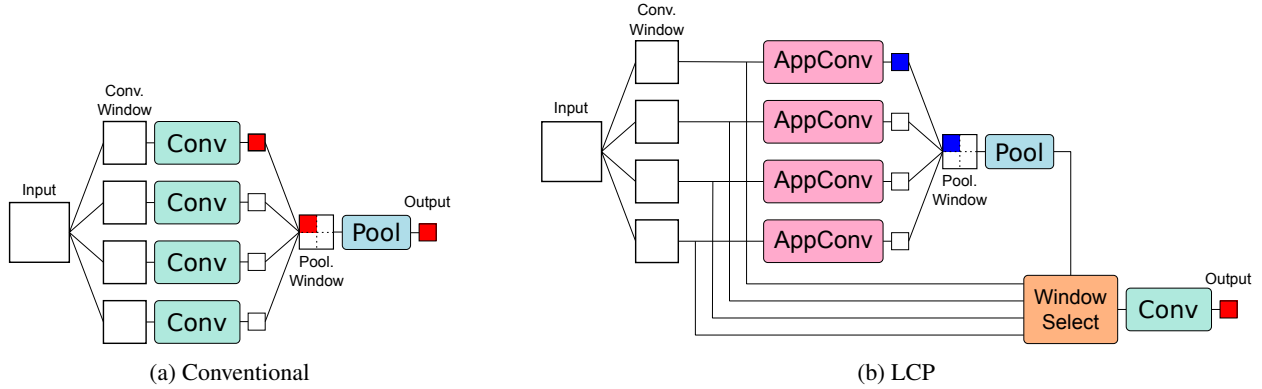


Figure 2: Convolution and pooling operations of the conventional method and the proposed method, LazyConvPool (LCP). (a) The conventional one executes exact convolutions (Conv) against all the windows. (b) LCP only executes an exact convolution (Conv) against the specified window predicted by the approximate convolution (AppConv).

architectures that implement CNN are proposed in [4, 5, 11, 13]. nn-X [11] is a salable co-processor for CNN. It consists of an array of possessing elements, which are dedicated hardware to accelerate the main components of CNN (convolution, pooling, and activation). HWCE [5] is also an accelerator for CNN particularly focusing on 2D convolutions, which are the most computationally intensive part in CNN. Both of the architectures intend to speed up the calculation of convolutions in a highly parallel manner by adopting an array structure and pipelined streaming processing units.

Another approach for low-power CNN processors is based on *approximate computing*, which is a computing method that aggressively uses approximation for the computation that does not require exact results. This is useful for the applications such as image processing and recognition tasks that tolerate some errors in the output. ApproxANN [23] and AxNN [21] are examples of using the approximate computing with neural networks. However, it is difficult to simply apply the approximate computing to CNN without spoiling recognition accuracy because of the deep structure of the CNN, where the approximation errors are accumulated and inflated during the propagation.

In this work, we propose a novel method to enhance the energy efficiency of the existing CNN processor like nn-X and HWCE by utilizing an approximation technique. In order to suppress the degradation of recognition performance, we adopt the approximation only to predict the region to be selected in the max pooling. With this approach, we can effectively eliminate redundant computation according to the prediction results calculated with a small overhead.

### 3. Proposed method

In this section, we propose LazyConvPool (LCP), an approximate computational method that combines convolution and max pooling for reducing power consumption of CNN processors. We also propose Sign Connect, a simplified arithmetic unit that can be effectively used with LCP.

#### 3.1. LazyConvPool

The objective of LCP is to eliminate the convolutional computation that will not affect the output of the succeeding pooling layer. Figure 2a illustrates the conventional convolution and pooling operations. In this example, four convolutions are first executed on an input image to calculate adjacent four feature values in a  $2 \times 2$  pooling window. Then

the max pooling is operated to pick up one representative feature that has a maximum value (shown as a red pixel). Worthy of attention here is that only one convolution result is passed to the next stage and all the other convolution results are discarded by the pooling operation. This means that the non-representative feature values do not affect the feature map after pooling as long as the feature selection in the pooling is correct.

On the basis of this observation, we construct LCP framework with two main parts: (1) approximated convolution units (AppConv) to predict the feature selection in the pooling layer, and (2) a single regular (and exact) convolution unit (Conv) to perform convolution only on the predicted window. Figure 2b illustrates the LCP-based convolution and pooling operations. The computation flow of LCP is summarized as follows:

- Step 1** Compute AppConv using the lightweight and approximate computation.
- Step 2** Operate max pooling to the results of AppConv to specify the convolution index of which the convolution result is propagated.
- Step 3** Compute Conv only to the input corresponding to the specified index in Step 2.

LCP can reduce the number of the exact convolutions by  $1/N_P$ , where  $N_P$  is the number of the features in a pooling window. On the other hand, the prediction by AppConv introduces additional hardware. To minimize additional power consumption associated with the prediction, we utilize lightweight and approximate computation called Sign Connect, which is described in the next subsection.

### 3.2. Sign Connect

LCP executed only the necessary convolution by computing AppConv before Conv. Computational cost of AppConv must be low to realize power reduction through LCP. Here, for this purpose, we propose Sign Connect as an approximate computation unit of reduced computation complexity.

Sign Connect is equivalent to “sign function” used in BinaryNet [7]. Sign Connect computes product terms by the following equations:

$$\begin{cases} P(W(k,l) = 1) &= 1 \\ P(W(k,l) = 0) &= 0 \end{cases}, \quad (8)$$

$$\begin{cases} P(W(k,l) = -1) &= 1 \\ P(W(k,l) = 0) &= 0 \end{cases}. \quad (9)$$

Contrary to Ternary Connect [19], the proposed equations are deterministic, thus random number generators are not necessary. Sign Connect can be implemented so that general multipliers are replaced by the multiplications of the

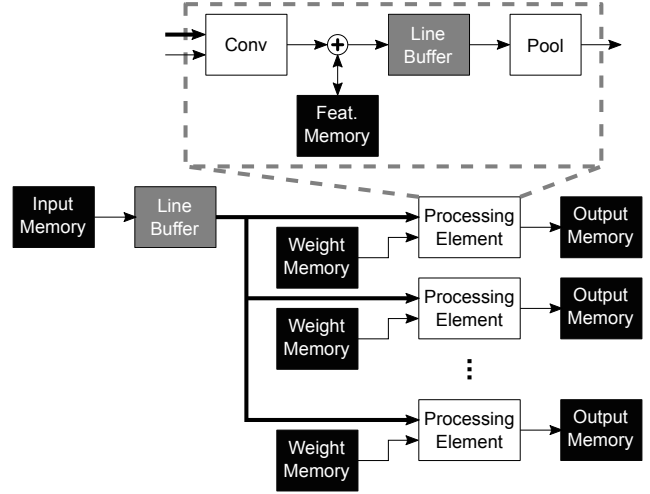


Figure 3: Base architecture for CNN processor.

sign of weights, which can be typically realized by multiplexers. With Sign Connect, multiplication could be drastically simplified, and the total performance becomes deterministic.

## 4. Hardware architecture

### 4.1. Base architecture

Figure 3 shows a block diagram of the base architecture of CNN processor. This is similar to those of nn-X [11] and HWCE [5], employing a parallel architecture with a shared memory and multiple processing elements. Each processing element, connected to the shared memory, executes convolution, pooling, and activation. In this hardware architecture, each element computes individual feature maps on the basis of a common input image (or feature map). The input image is supplied to each element via a line buffer. The line buffer enables a continuous and efficient data feed for convolutions on a sliding windows. Weights of the convolution layers are supplied from a local memory and kept in registers in convolution circuits during computation. At the following stage of convolution circuits, the memory and accumulators are connected to accumulate the value of feature maps. Convolution results are accumulated on the memory that has the size of the feature map.

### 4.2. LCP architecture

Figure 4 shows the proposed processing element architecture in which LCP is applied to the base architecture. Buffered inputs are supplied to AppConv and Conv with weights used for convolution. A processing element first computes AppConv and pooling. The results are temporarily stored in the index memory connected to the pooling circuit and utilized as the signal that control inputs to the Conv

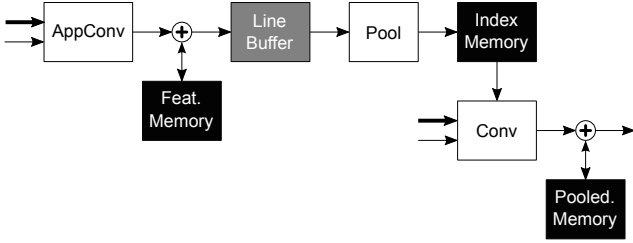


Figure 4: The proposed architecture for LCP-based CNN processor.

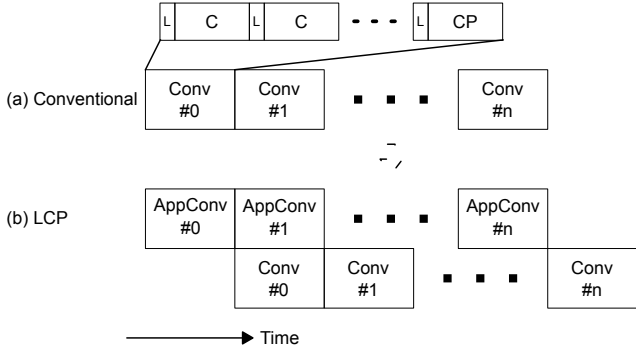


Figure 5: Timing chart of CNN calculation by (a) the conventional and (b) the proposed LCP methods.

circuit. Conv results are accumulated only for the size of the pooled feature map.

Figure 5 shows the timing charts of CNN computation with and without LCP. As shown in Figure 5a, the conventional CNN on the base architecture repeatedly executes sets of convolutions. Each set consists of weight loading (L) and convolution (C). In the final set, pooling is executed just after the convolutions (CP). As for LCP, the two kinds of convolutions (AppConv and Conv) must be processed and this may extend the execution time. In order to minimize the additional execution time caused by LCP, the proposed architecture conducts the two kinds of convolutions in a pipelined way as shown in Figure 5b. Thus, the overhead of execution time is reduced to just one computing routine time.

## 5. Experiments

### 5.1. Classification accuracy

**Datasets** We evaluated the classification accuracy of LCP with image recognition datasets. We used two datasets: the handwritten digits dataset MNIST [18] and the object recognition dataset CIFAR-10 [16]. For MNIST, we preprocessed the images by the normalization to limit the range of data values to 0–1. For CIFAR-10, we preprocessed images by a) global contrast normalization, b) ZCA whitening [16], and c) horizontal flips. a) and b) are preprocessing

| Layer name      | Parameter                |
|-----------------|--------------------------|
| Input image     | size: 28x28, channel: 1  |
| Convolution     | kernel: 5x5, channel: 20 |
| Max pooling     | kernel: 2x2, stride: 2   |
| ReLU            |                          |
| Convolution     | kernel: 5x5, channel: 50 |
| Max pooling     | kernel: 2x2, stride: 2   |
| ReLU            |                          |
| Fully connected | channel: 500             |
| ReLU            |                          |
| Fully connected | channel: 10              |
| Softmax         |                          |

Table 1: Network parameters of CNN for MNIST dataset.

| Layer name      | Parameter                             |
|-----------------|---------------------------------------|
| Input image     | size: 32x32, channel: 3               |
| Convolution     | kernel: 5x5, channel: 64, padding: 2  |
| Max pooling     | kernel: 2x2, stride: 2                |
| ReLU            |                                       |
| Convolution     | kernel: 5x5, channel: 128, padding: 2 |
| Max pooling     | kernel: 2x2, stride: 2                |
| ReLU            |                                       |
| Convolution     | kernel: 5x5, channel: 256, padding: 2 |
| Max pooling     | kernel: 2x2, stride: 2                |
| ReLU            |                                       |
| Convolution     | kernel: 5x5, channel: 512, padding: 2 |
| Max pooling     | kernel: 2x2, stride: 2                |
| ReLU            |                                       |
| Fully connected | channel: 512                          |
| ReLU            |                                       |
| Fully connected | channel: 10                           |
| Softmax         |                                       |

Table 2: Network parameters of CNN for CIFAR-10 dataset.

techniques for normalization. c) is applied for data augmentation without changing the image size. We trained networks by all data given in the training set and evaluated target methods by 100 data in the test set none of which exists in the training set. Network parameters are shown in Tables 1 and 2.

**Results** Classification results are shown in Table 3. We evaluated the four types of CNN: (a) the conventional CNN, (b) CNN with LCP that computes AppConv by the exact convolution, (c) AppConv by Ternary Connect, and (d) AppConv by Sign Connect. The results of LCP with Ternary

| Method           | MNIST | CIFAR-10 |
|------------------|-------|----------|
| (a) Conventional | 0.986 | 0.860    |
| (b) LCP          | 0.986 | 0.860    |
| (c) LCP + TC     | 0.984 | 0.825    |
| (d) LCP + SC     | 0.988 | 0.849    |

Table 3: Classification accuracy on datasets.

Connect are the average of 10 independent runs because of the probabilistic nature of the Ternary Connect.

The results for LCP with exact convolution is completely equal to the conventional one because this is equivalent to the addition of redundant operation. For LCP with Ternary Connect, the accuracy is almost the same in the experiment of MNIST. However, the accuracy degrades about 3.5% in CIFAR-10. The difference shows the performance degradation due to the prediction inaccuracy using Ternary Connect for the complex data such as color images. For LCP with Sign Connect, the accuracy is almost equal to the conventional result in MNIST data set. Regarding to CIFAR-10, the accuracy degradation has been reduced compared to LCP with Ternary Connect.

## 5.2. Power consumption

**Experimental setup** To evaluate power consumption, we implemented both of the base architecture and the proposed LCP architecture with hardware description language (HDL). The dynamic power consumption is evaluated by using the netlists synthesized from the HDL and the logic simulation results that annotate the switching activity in the netlists. We used two commercial logic cell libraries of 65 nm process at 0.55 V and 28 nm process at 0.92 V.

In the power estimation, we focused on a mid-level layer when recognizing 10 MNIST samples. The numbers of input and output feature maps for this layer are 20 and 50, respectively. CNN processor implements  $P = 4$  processing elements operating at a 100 MHz clock frequency.

**Results** Figures 6 and 7 show the average power consumption for each digit in the base and the proposed LCP architecture. We can see that the power consumption of the LCP architecture decreases for all digits.

Figures 8 and 9 show the average power consumption for all digits in the base and the proposed LCP architectures. The breakdowns of power consumption for major modules are also presented. It can be seen that the power consumption in Conv modules is greatly reduced by using LCP. Although the power consumption in AppConv and the other modules is increased, the total power decreases by 20.2% and 17.8% for 65 nm and 28 nm process libraries, respectively. Considering the increase of the execution time discussed in 4.2, which is 7.7% in this experiment, the en-

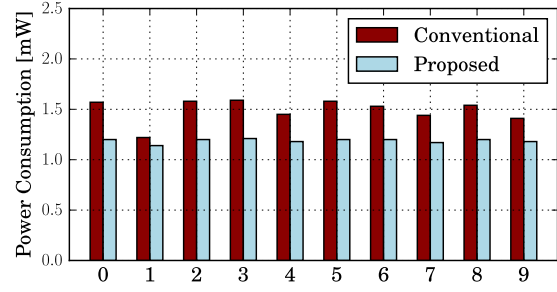


Figure 6: Power consumption for classifying each digit in the CNN architectures with 65 nm process library.

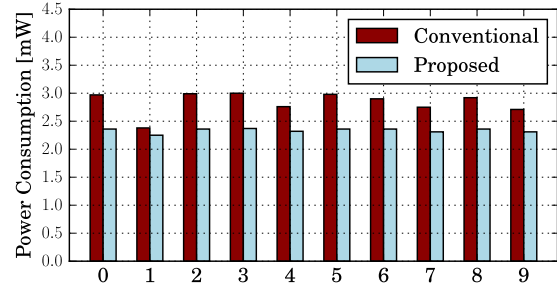


Figure 7: Power consumption for classifying each digit in the CNN architectures with 28 nm process library.

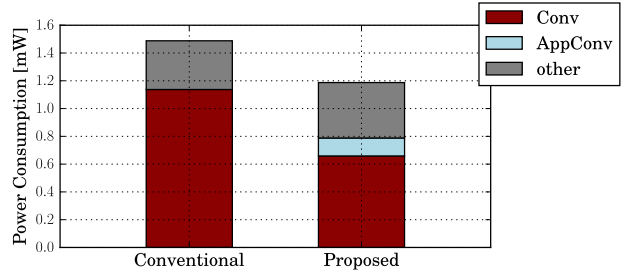


Figure 8: Average power consumption and breakdowns for the major modules in the CNN architectures with 65 nm process library.

ergy reduction by LCP is 14.1% and 11.4% for 65 nm and 28 nm process libraries, respectively.

## 6. Conclusion

In this paper, a new approach to reduce the number of multiplications in CNN based on an approximated prediction is proposed. Through the experimental evaluation, the proposed LazyConvPool algorithm and its hardware architecture successfully reduce the energy consumption for image recognition by 11.4%–14.1% while keeping the recognition accuracy. The reduction rate is limited in this pa-

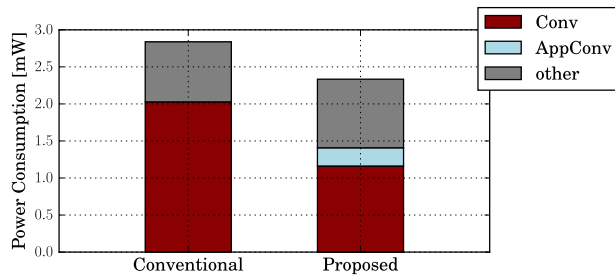


Figure 9: Average power consumption and breakdowns for the major modules in the CNN architectures with 28 nm process library.

per because we only focused on the pooling operations. For a future work, we will apply the proposed approach to other kinds of selective operations in neural networks such as Maxout techniques [12] and network compression techniques [3, 14] to realize more energy-efficient CNN processors.

## References

- [1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1
- [2] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen. Compressing convolutional neural networks. *Computing Research Repository*, abs/1506.04449, 2015. 2
- [3] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. In *Proceedings of International Conference on Machine Learning*, pages 2285–2294, 2015. 2, 7
- [4] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. In *Proc. IEEE International Solid-State Circuits Conference*, pages 262–263, 2016. 3
- [5] F. Conti and L. Benini. A ultra-low-energy convolution engine for fast brain-inspired vision in multicore clusters. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pages 683–688, 2015. 3, 4
- [6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 1
- [7] M. Courbariaux and Y. Bengio. Binarynet: Training deep neural networks with weights and activations constrained to +1 or −1. *Computing Research Repository*, abs/1602.02830, 2016. 2, 4
- [8] N. Dalai, B. Triggs, I. Rhone-Alps, and F. Montbonnot. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 1
- [9] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 1
- [10] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1
- [11] V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello. A 240 G-ops/s mobile coprocessor for deep neural networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pages 696–701, 2014. 3, 4
- [12] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the International Conference on Machine Learning*, pages 1319–1327, 2013. 7
- [13] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally. EIE: efficient inference engine on compressed deep neural network. *Computing Research Repository*, abs/1602.01528, 2016. 3
- [14] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *Computing Research Repository*, abs/1510.00149, 2015. 2, 7
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Computing Research Repository*, abs/1512.03385, 2015. 1
- [16] A. Krizhevsky. Learning multiple layers of features from tiny images. Master’s thesis, University of Toronto, 2009. 5
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Neural Information Processing Systems*, pages 1097–1105, 2012. 1
- [18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5
- [19] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio. Neural networks with few multiplications. *Computing Research Repository*, abs/1510.03009, 2015. 2, 4
- [20] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999. 1
- [21] S. Venkataramani, A. Ranjan, K. Roy, and A. Raghunathan. AxNN: Energy-efficient neuromorphic systems using approximate computing. In *Proceedings of the International Symposium on Low Power Electronics and Design*, pages 27–32, 2014. 1, 3
- [22] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 1
- [23] Q. Zhang, T. Wang, Y. Tian, F. Yuan, and Q. Xu. Approx-ANN: An approximate computing framework for artificial neural network. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pages 701–706, 2015. 1, 3

# **APPENDIX UJIIIE02**



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

ABOUT BLOG PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER PEOPLE

INTERNET ARCHIVE

DONATE

Wayback Machine

Explore more than 946 billion web pages saved over time

http://www.cv-foundation.org/openaccess/CVPR2016\_workshop:

Calendar

Collections

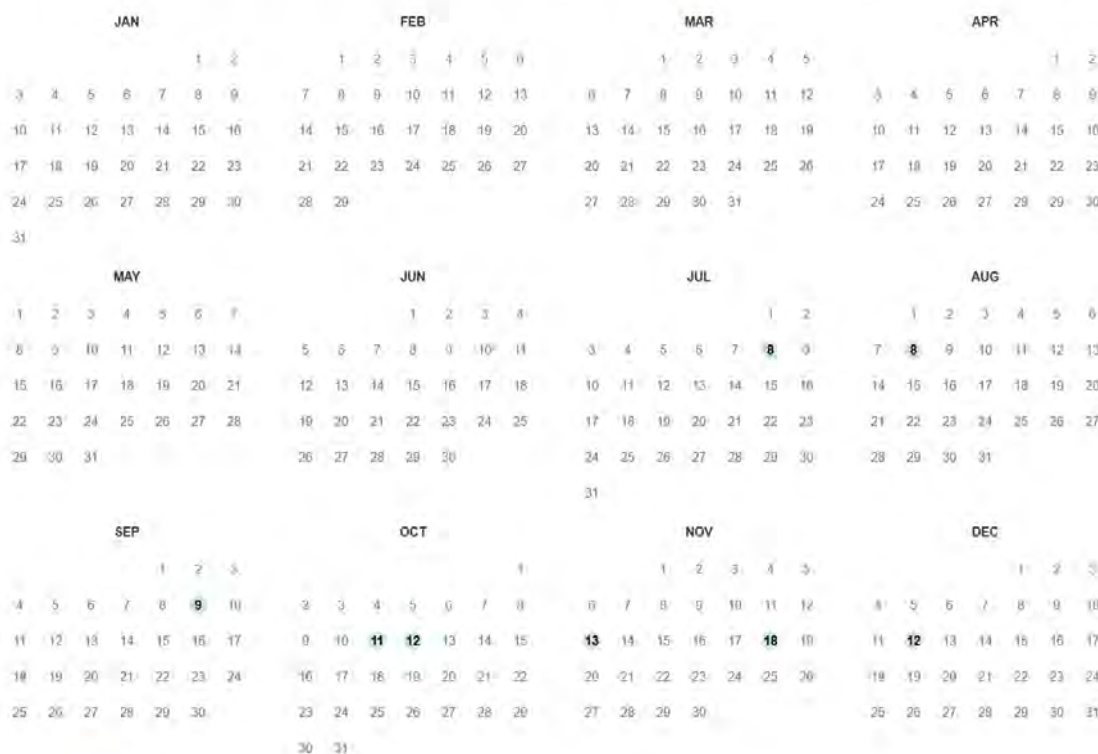
Changes

Summary

Site Map

URLs

Saved 39 times between July 8, 2016 and January 24, 2025.



### Note

This calendar view maps the number of times [http://www.cv-foundation.org/openaccess/CVPR2016\\_workshops/menu.py](http://www.cv-foundation.org/openaccess/CVPR2016_workshops/menu.py) was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the [FAQ](#).

[FAQ](#) | [Contact Us](#) | [Terms of Service](#) (Dec 31, 2014)



The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's [Terms of Use](#).



## Computer Vision Foundation open access

These research papers are the Open Access versions, provided by the [Computer Vision Foundation](#).  
The authoritative versions of these papers are posted on [IEEE Xplore](#).

*This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright.*

### CVPR 2016 Workshops, Las Vegas Nevada

[Computer Vision in Vehicle Technology](#)

[Biometrics](#)

[Perception Beyond the Visible Spectrum](#)

[DeepVision: Deep Learning in Computer Vision](#)

[Egocentric Vision](#)

[Embedded Vision](#)

[Biomedical Image Registration](#)

[Computational Cameras and Displays](#)

[Large Scale 3D Data: Acquisition, Modelling and Analysis](#)

[ChaLearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge](#)

[Visual Analysis of Satellite to Street Imagery](#)

[Performance Evaluation of Tracking and Surveillance](#)

[Observing and Understanding Hands in Action](#)

[Differential Geometry in Computer Vision and Machine Learning](#)

[Robust Features for Computer Vision](#)

[Automatic Traffic Surveillance](#)

[Computer Vision for Microscopy Image Analysis](#)

[Context-Based Affect Recognition and Affective Face In-the-Wild](#)

[Moving Cameras Meet Video Surveillance: From Body-Borne Cameras to Drones](#)

[Back](#)

# **APPENDIX UJIIIE03**



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

ABOUT BLOG PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER PEOPLE

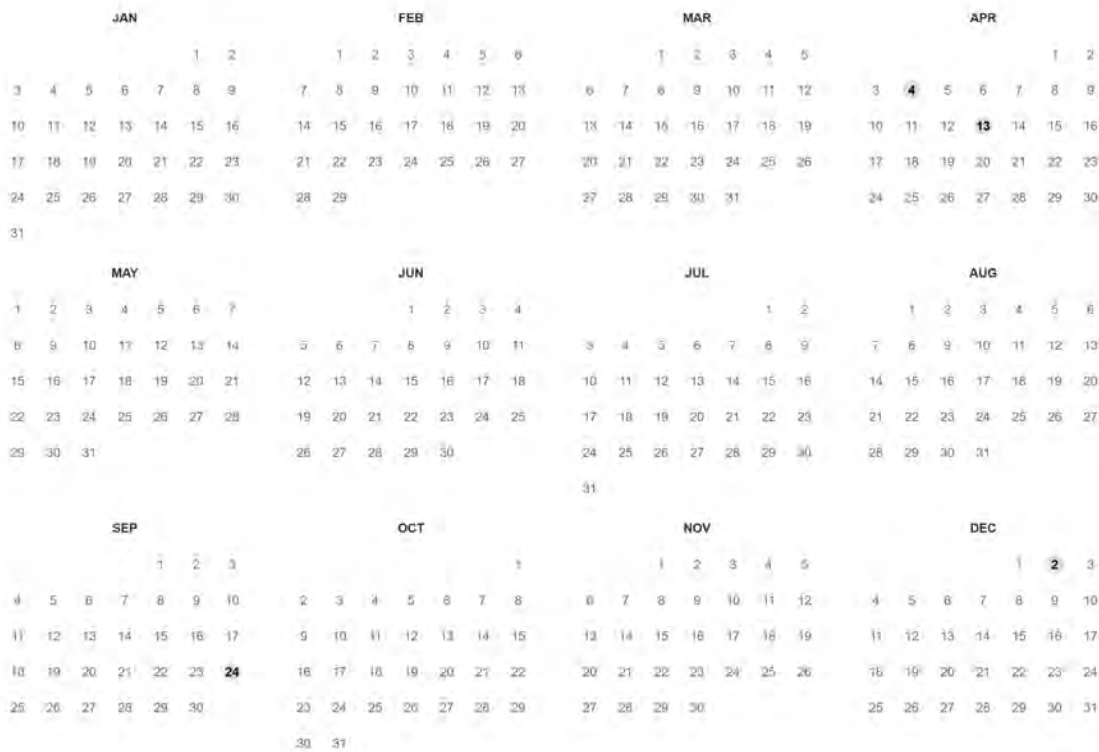


Explore more than 946 billion web pages saved over time

http://cvisioncentral.com/promotion/evw2016/

Calendar Collections Changes Summary Site Map URLs

Saved 9 times between April 4, 2016 and July 30, 2024.



### Note

This calendar view maps the number of times <http://cvisioncentral.com/promotion/evw2016/> was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the [FAQ](#).

[FAQ](#) | [Contact Us](#) | [Terms of Service](#) (Dec 31, 2014)



The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & [Archive-It.org](#).

Your use of the Wayback Machine is subject to the Internet Archive's [Terms of Use](#).

Computer Vision Central

- Home
- Activity
- Jobs
- Blogs
- Community
- Resource
- Labs
- Conference



Feb 4, 2016 at 1:54 am in by Computer Vision Central

Email or

Share

The Twelfth IEEE Embedded Vision Workshop

July 1st, 2016,  
Las Vegas, NV, USA  
Held in conjunction with IEEE CVPR 2016.  
[www.cvvisioncentral.com/evw2016](http://www.cvvisioncentral.com/evw2016)  
Call for Papers: (pdf)

Program

Location: Emperor II  
Schedule: Full day

| Start Time | Paper/Talk Title                                                                                               | Author/Speaker                                                                                                                                                                                                                              |
|------------|----------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 8:00       | Welcome                                                                                                        |                                                                                                                                                                                                                                             |
| 8:10       | Keynote: Vision-Centric ADAS Silicon Architectures for Platform Efficiency, Intelligence and Flexibility       | Chris Rowen (Cadence IP, USA)                                                                                                                                                                                                               |
| 8:55       | A Diverse Low Cost High Performance Platform for Advanced Driver Assistance System (ADAS) Applications         | Prashanth Viswanath, Pramod Swami, Mihir Mody, Kedar Chitnis, Sujith Shivalingappa, Soyeb Nagori, Manu Mathew, Kumar Desappan, Shyam Jagannathan, Deepak Poddar, Anshu Jain, Hrushikesh Garud, Vikram Appia, Mayank Mangla, Shashank Dabral |
| 9:15       | A Visual Attention Algorithm Designed for Coupled Oscillator Acceleration                                      | Christopher Thomas, Adriana Kovashka, Donald Chiarulli, Steven Levitan                                                                                                                                                                      |
| 9:35       | Embedded Motion Detection via Neural Response Mixture Background Modeling                                      | Mohammad Javad Shafiee, Parthipan Siva, Paul Fieguth, Alexander Wong                                                                                                                                                                        |
| 10:00      | Morning Break                                                                                                  |                                                                                                                                                                                                                                             |
| 10:30      | Keynote: The Challenges and Fun of Building Mu, a Tiny Smart Eye for Toys                                      | Tianli Yu (Morpx Inc., USA)                                                                                                                                                                                                                 |
| 11:15      | A Scalable High-Performance Hardware Architecture for Real-Time Stereo Vision by Semi-Global Matching          | Jaco Hofmann, Andreas Koch, Jens Korinth                                                                                                                                                                                                    |
| 11:35      | Vision Based Autonomous Orientational Control for Aerial Manipulation via on-board FPGA                        | Suphachart Leewiwatwong, Shouhei Shimahara, Robert Ladig, Kazuhiro Shimonomura                                                                                                                                                              |
| 11:55      | Embedded Vision System for Atmospheric Turbulence Mitigation                                                   | Ajinkya Deshmukh, Gaurav Bhosale, Swarup Shanti, Karthik Reddy, Hemanthkumar P. Chandrasekhar A, Kirankumar P, Vijaysagar K                                                                                                                 |
| 12:35      | Lunch (on your own)                                                                                            |                                                                                                                                                                                                                                             |
| 13:30      | Keynote: Energy-Efficient Hardware for Embedded Vision and Deep Convolutional Neural Networks                  | Vivienne Sze (MIT, USA)                                                                                                                                                                                                                     |
| 14:15      | Spotlights Posters and Demos                                                                                   |                                                                                                                                                                                                                                             |
| 14:35      | Posters                                                                                                        |                                                                                                                                                                                                                                             |
|            | Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor      | Takayuki Ujiie, Masayuki Hiromoto, Takashi Sato                                                                                                                                                                                             |
|            | Visual Monocular Obstacle Avoidance for Small Unmanned Vehicles                                                | Levente Kovacs                                                                                                                                                                                                                              |
|            | Real-time, Embedded Scene Invariant Crowd Counting Using Scale-Normalized Histogram of Moving Gradients (HoMG) | Parthipan Siva, Mohammad Javad Shafiee, Michael Jamieson, Alexander Wong                                                                                                                                                                    |
|            | 3DCapture: 3D Reconstruction for a Smartphone                                                                  | Oleg Muratov, Yury Slynko, Vitaly Chernov, Maria Lyubimtseva, Artem Shamsuarov, Victor Bucha                                                                                                                                                |
|            | Embedded Computing Framework for Vision-based Real-Time Surround Threat Analysis and Driver Assistance         | Frankie Lu, Sean Lee, Ravi Kumar Satzoda, Mohan Trivedi                                                                                                                                                                                     |
| 15:25      | Afternoon Break                                                                                                |                                                                                                                                                                                                                                             |
| 16:00      | Keynote: Multi-Sensor Fusion for Self-Healing Vision Based Navigation                                          | Stephan Weiss (Alpen-Adria-Universität, Austria)                                                                                                                                                                                            |
| 16:45      | Demos                                                                                                          |                                                                                                                                                                                                                                             |
|            | Aquifi SIR: Extremely Low Latency Depth and Color Camera with Computation on FPGA                              | Paolo Di Febbo, Alessandro Mattei, Alessandro Temil, Aryan Hazezghi, Stefano                                                                                                                                                                |

|       |                                                                                                              |                                                   |
|-------|--------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
|       | Embedded Computing Framework for Vision-based Real-Time Surround Threat Analysis and Driver Assistance       | Abbas Rafii, Jason Trachewsky                     |
|       | Embedding Convolutional Neural Networks for Real-Time Vehicle Localization Into Consumer Grade Mobile Phones | Frankie Lu, Sean Lee, Ravi Satzoda, Mohan Trivedi |
| 17:30 | Closing Remarks and Awards                                                                                   | Nauto, Inc.                                       |

## Sponsors

Best paper awards are provided by the following sponsors:

- Morpx Inc.: Computer vision development kit for Mu Toys
- Mathworks: License to Matlab and additional toolboxes along with a Raspberry PI 3
- NVIDIA: GPU GeForce GTX Titan X.



## Call For Participation

Embedded Vision is what makes Computer Vision mainstream today, as it brings together embedded systems with vision functionalities. Due to the emergence of powerful yet low-cost and energy-efficient processors, it has become possible to incorporate vision capabilities into a wide range of embedded systems, including video search and annotation, surveillance, gesture recognition in video games, driver assist systems in automotive safety, and autonomous robots such as drones. The IEEE Embedded Vision Workshop (EVW) brings together researchers working on vision problems that share embedded system characteristics.

## Important Dates

- Paper submission: March 18 (05:00pm PST), 2016
- Notification to the authors: April 27, 2016
- Camera ready paper: May 2, 2016

Paper submission website: <https://cmt3.research.microsoft.com/EVW2016>

This year, we strongly encourage submissions on robotic systems and applications. Research papers are solicited in, but not limited to, the following topics:

- Deep learning hardware and software architectures for visual exploitation
- Analysis of vision problems specific to embedded systems
- Analysis of embedded systems issues specific to computer vision
- Large scale computer vision problems including object recognition, scene analysis, visual, industrial and medical applications
- Embedded vision for robotics (industrial, mobile, and consumer)
- Embedded vision for Unmanned Ground and Air Vehicles, including consumer drones
- Embedded visual navigation (mapping, localization, VSLAM, VO)
- New trends in programmable processors for vision and computational imaging
- Applications of and algorithms for embedded vision on:
  - Massively parallel platforms such as GPUs (PC, embedded and mobile)
  - Programmable platforms such as DSPs and multicore SoCs
  - Reconfigurable platforms such as FPGAs and SoCs with reconfigurable logic
  - Mobile devices, including smartphones and tablets
  - Vision-based client devices for the Internet of Things (IoT)
- Embedded vision for 3D movies and TV
- Biologically-inspired vision and embedded systems
- Computer vision applications distributed between embedded devices and servers
- Social networking embedded computer vision applications
- Educational methods for embedded vision
- User interface designs and CAD tools for embedded vision applications
- Hardware enhancements (lens, imager, processor) that impact vision applications
- Software enhancements (OS, middleware, vision libraries, development tools) that impact embedded vision application
- Methods for standardization and measurement of vision functionalities
- Performance metrics for evaluating embedded systems
- Hybrid embedded systems combining vision and other sensor modalities

## Demo Session

This year, EVW will have live demonstrations of embedded vision prototypes and solutions. During this session, authors, engineers, and researchers can showcase their prototypes with real-time implementations of vision systems on embedded computing platforms. Authors of selected papers (both oral and poster) will be invited to demonstrate their works. Additionally, we also invite abstracts independent of the paper submissions to present at the demo session.

IPR2025-01399

during the workshop. Selected abstracts will be given an opportunity to demonstrate their solutions during the workshop.

Please send your demo abstract/proposal to [rsatzoda@eng.ucsd.edu](mailto:rsatzoda@eng.ucsd.edu).

Deadline: **April 20, 2016**

## Special Journal Issue on Embedded Computer Vision

All of the previous Workshops on Embedded (Computer) Vision (ECVV and EVW) were held in conjunction with CVPR, with the exception of the fifth, which was held in conjunction with ICCV 2009. These events were very successful, and selected workshop papers have been published in two special issues of major journals (EURASIP Journal on Embedded Systems and CVIU) and in two Springer monographs titled Embedded Computer Vision.

This year, there will be a Special Journal Issue on Embedded Computer Vision within Springer's Journal of Signal Processing Systems. Authors of selected papers will be invited to submit extended versions of their workshop papers for publication in the special issue.

Special Journal Issue deadline: **Aug 14, 2016**

CFP

## Organizing Committee

### General Chair:

Stefano Mattoccia, University of Bologna, Italy  
Jagadeesh Sankaran, Nvidia

### Program Chair:

Martin Humenberger, AIT Austrian Institute of Technology  
Swarup Medasani, Urmil Systems

### Demo Chair:

Ravi Kumar Satzoda, UC San Diego

### Sponsorship Chair:

Goksel Dedeoglu, PercepTonic

### Steering Committee:

Ahmed Nabil Belbachir, Teknova AS (Norway)  
Sek Chai, SRI International  
Margrit Gelautz, Vienna University of Technology  
Branislav Kisanin, Nvidia  
Fridtjof Stein, Daimler AG

### Web Chair:

Yu Wang, Ambarella

### Program Committee (tentative):

Abelardo Lopez-Lagunas, ITESM  
Antonio Haro, Navteq  
Bernhard Rinner, Klagenfurt University  
Burak Ozer, Verificon  
David Ilstrup, Honda Research Institute  
David Moloney, Movidius  
Dipan Mandal, Intel Corporation  
Hassan Rabah, University of Lorraine  
Hongying Meng, Brunel University  
Khanh Duc, Nvidia  
Kofi Appiah, Nottingham Trent University  
Linda Wills, Georgia Tech  
Marilyn Wolf, Georgia Tech  
Peter Venetianer, Digital Signa Corp.  
Rajesh Narasimha, MetaIO  
Sebastiano Battiato, U. of Catania  
Steve Mann, University of Toronto  
Sven Fleck, SmartSurv  
Terry Boult, U. of Colorado  
Vinay Sharma, Apple  
Vittorio Murino, Istituto Ital. di Tecn.  
Zhu Li, Hong Kong Polytechnic University  
Senyo Apewokin, Texas Instruments  
Moshe Ben-Ezra, Microsoft  
Faycal Bensaali, Qatar University

Xin Chen, Naivety  
Rita Cucchiara, U. of Modena and Reggio Emilia  
Orazio Gallo, Nvidia  
Masatoshi Ishikawa, U. of Tokyo  
Kihwan Kim, Nvidia  
Kevin Koeser, ETH Zurich  
Zhu Li, Hong Kong Polytechnic U.  
Darnell Moore, Texas Instruments  
Andre Morin, Lyrtech  
Zoran Nikolic, Texas Instruments  
Sankalita Saha, NASA  
Mainak Sen, Cisco Systems  
Dabral Shashank, Texas Instruments  
Salvatore Vitabile, U. of Palermo  
Ruigang Yang, U. of Kentucky  
Witek Jachimczyk, Mathworks  
Zoran Zivkovic, Intel Corporation  
Stephan Weiss, Klagenfurt University  
Roland Brockers, JPL  
Thomas Kadiofsky, AIT Austrian Institute of Technology  
Markus Murschitz, AIT Austrian Institute of Technology  
Nicolas Thorstensen, IVISO GMBH

## Previous Embedded Computer Vision Workshops (ECVW)

- [ECVW2015](#)
- [ECVW2014](#)
- [ECVW2013](#)
- [ECVW2012](#)
- [ECVW2011](#)
- [ECVW2010](#)
- [ECVW09](#)
- [ECVW08](#)
- [ECVW07](#)
- [ECVW06](#)
- [ECVW05](#)

← VIEW 2016: Fifth Annual Vision Industry and Entrepreneur Workshop (Held in conjunction with CVPR 2016)

[Next Post](#) →

Comments are closed.

[About Us](#) | [Contact Us](#) | [Privacy Policy](#) | [Terms of Use](#)

Copyright © 2008-2016 Computer Vision Central. All rights reserved.

# APPENDIX UJIIIE04



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

ABOUT BLOG PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER PEOPLE

INTERNET ARCHIVE

DONATE

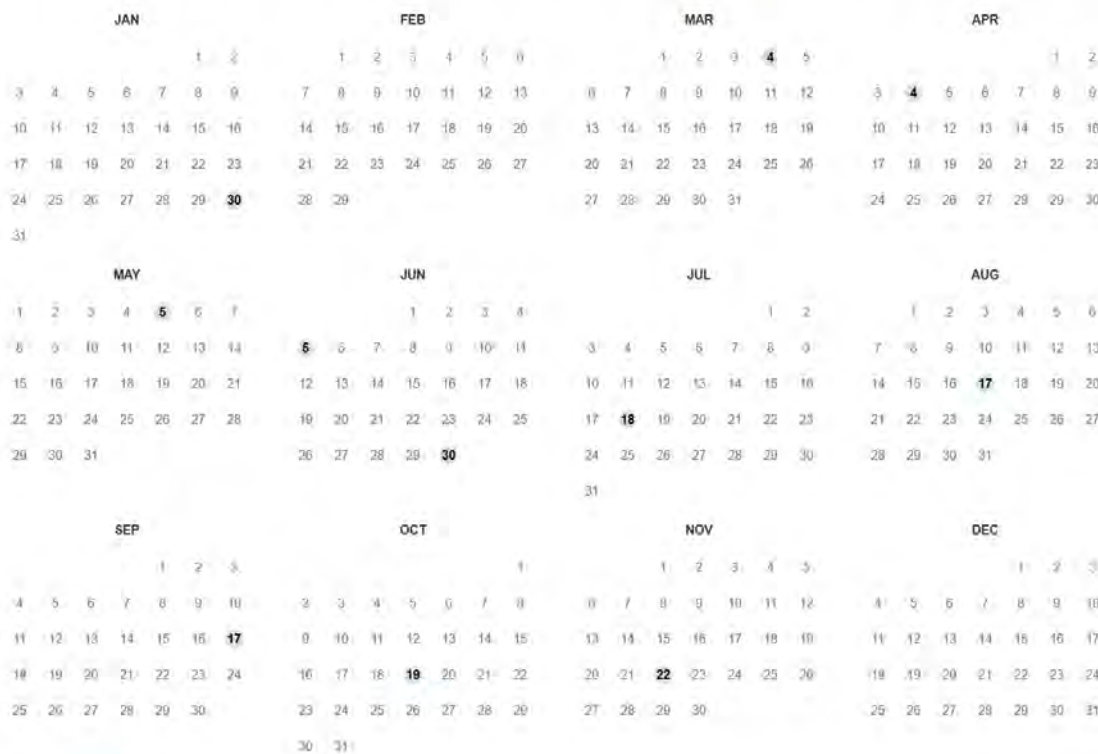
WayBackMachine

Explore more than 946 billion web pages saved over time

http://cvpr2016.thecvf.com/program/industry\_expo

Calendar Collections Changes Summary Site Map URLs

Saved 40 times between January 30, 2016 and August 9, 2025.



### Note

This calendar view maps the number of times [http://cvpr2016.thecvf.com/program/industry\\_expo](http://cvpr2016.thecvf.com/program/industry_expo) was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the [FAQ](#).

[FAQ](#) | [Contact Us](#) | [Terms of Service](#) (Dec 31, 2014)



The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive.org.

Your use of the Wayback Machine is subject to the Internet Archive's [Terms of Use](#).

INDUSTRY EXPO

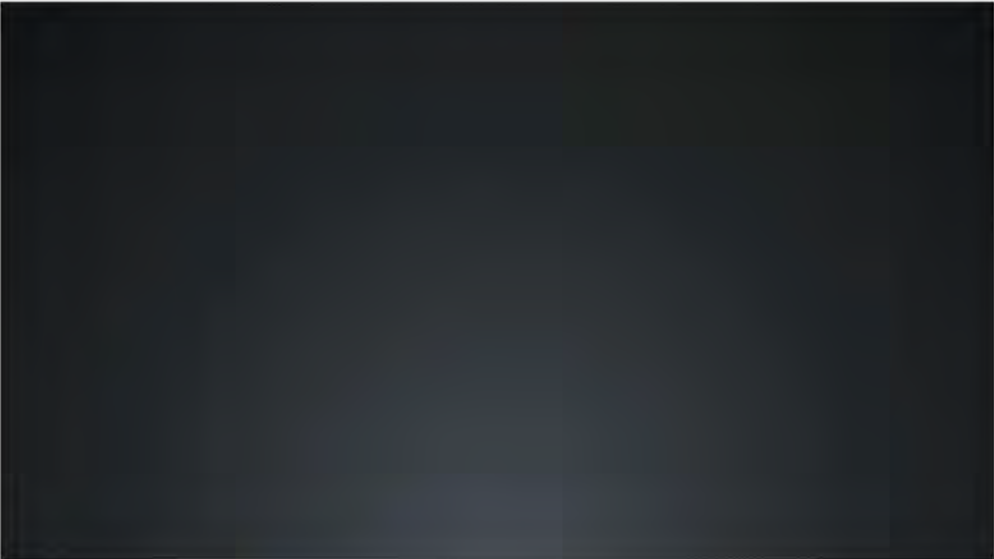
SPONSORS

EXHIBITOR/SPONSOR INFO

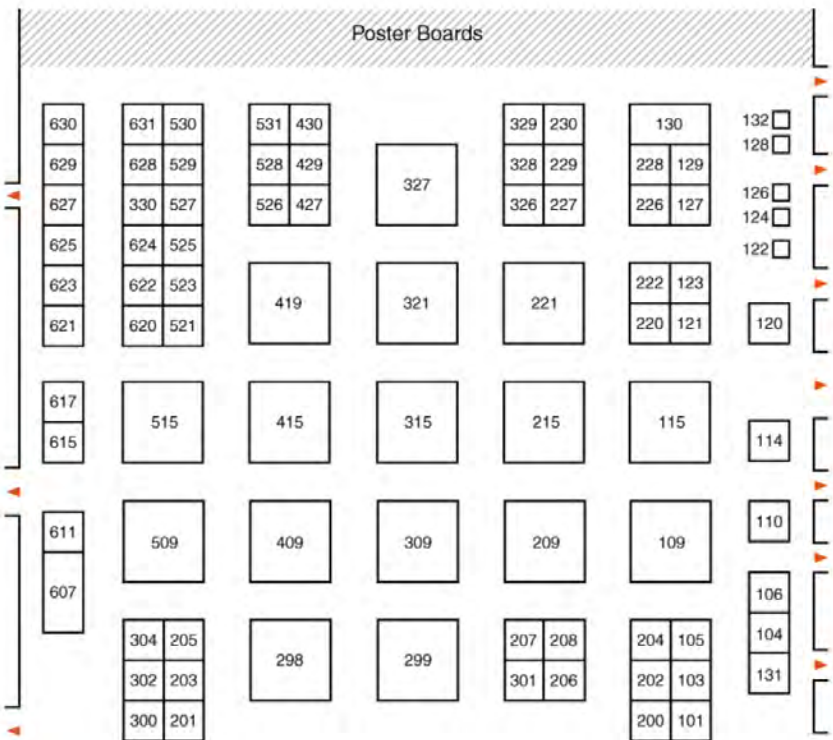
PAYMENT OPTIONS

CVPR 2016 Industry Expo

Computer vision innovation has created an exciting opportunity for a ground-breaking EXPO 2016. From promising startups and creative standouts to the biggest industry leaders, this expo will be a showcase of companies that are pushing the boundaries of what's possible with computer vision and machine learning technologies. For the first time in CVPR's history, we are providing an exciting, "trade-show" like atmosphere to foster maximal visibility and exposure for each on-site exhibitor. It is a unique opportunity for, academians, students, budding entrepreneurs, technologists and others to cross paths and exchange the latest and newest ideas. CVPR EXPO 2016 will run at Caesar's Palace in Las Vegas for the duration of IEEE CVPR, co-locating with the premier academic and technical presentations. Over 3,000 CVPR attendees will create a one-of-a-kind opportunity for networking, recruiting, inspiration and motivation.



EXPO is located in the Octavius Ballroom  
Open during the main conference: June 27-30





| Booth | Exhibitor                                       |
|-------|-------------------------------------------------|
| 531   | 3dMD                                            |
| 298   | a-eye                                           |
| 215   | A9                                              |
| 301   | Adobe Systems                                   |
| 115   | Amazon                                          |
| 509   | Apple                                           |
| 201   | Avigilon                                        |
| 528   | Baidu                                           |
| 523   | Beijing DeepGlint Information Technologies, Inc |
| 627   | blippar                                         |
| 622   | Bodylabs                                        |
| 526   | Bosch                                           |
| 430   | CEVA                                            |
| 126   | Cizr                                            |
| 429   | Cognex                                          |
| 631   | CSIRO DATA61                                    |
| 427   | DAQRI                                           |
| 200   | Datatang                                        |
| 521   | Dextro                                          |
| 515   | Didi                                            |
| 110   | Disney Research                                 |
| 409   | DJI                                             |
| 328   | Ebay                                            |
| 611   | Elsevier                                        |
| 415   | Face++                                          |
| 321   | Facebook                                        |
| 106   | FiveFocal                                       |
| 128   | GemHunt.com                                     |
| 327   | Google                                          |
| 326   | HERE                                            |
| 120   | HiScene                                         |
| 229   | IBM                                             |
| 529   | iniLabs                                         |
| 209   | Intel                                           |
| 530   | IntelliVision                                   |
| 227   | iRobot                                          |
| 302   | Kitware                                         |
| 132   | Lighthouse                                      |
| 228   | Linkface                                        |
| 628   | Magic Leap                                      |
| 104   | Markable                                        |
| 620   | Mathworks                                       |

|         |                                  |
|---------|----------------------------------|
| 230     | Matterport                       |
| 109     | MERL                             |
| 315     | Microsoft Research               |
| 615     | Morgan & Claypool                |
| 122     | Morpx                            |
| 623     | Movidijs                         |
| 630     | Nervana                          |
| 203     | NREC                             |
| 205     | Nuctech                          |
| 221     | NVIDIA                           |
| 617     | Occam Vision Group               |
| 625     | Omron                            |
| 105     | Panasonic                        |
| 629     | Pinterest                        |
| 130     | Point Grey                       |
| 309     | Prism                            |
| 114     | Qihoo                            |
| 124     | QR Hub Ltd                       |
| 207     | Samasource                       |
| 621     | Second Spectrum                  |
| 226     | SenseTime                        |
| 127     | SightHound                       |
| 222     | Skydio                           |
| 220     | Snapchat                         |
| 304     | Sony Electronic                  |
| 123     | Spare5                           |
| 131     | SportLogiq Inc.                  |
| 527     | Spotscale                        |
| 609+607 | Springer                         |
| 204     | SRI International                |
| 101     | Tencent                          |
| 206     | Tesla                            |
| 299     | Toyota Research Institute        |
| 419     | Tusimple                         |
| 129     | Twitter                          |
| 121     | Uber ATC                         |
| 525     | UISEE Co, Ltd                    |
| 103     | Umbo CV                          |
| 624     | uSens                            |
| 208     | Van Gogh Imaging                 |
| 330     | Visual Computing Center at KAUST |
| 329     | Wolfram Research                 |
| 202     | wmch inc.                        |
| 300     | Zoox                             |



© 2016 CVPR

# **APPENDIX UJIIIE05**



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue

# MARC View

LEADER 02331nas a2200457 a 4500

001 48647063

003 SIRSI

005 20250814090058.0

006 m o d

007 cr n

008 110331c200u9999njuar ss | a0eng

010 a| 2011204049

019 a| DO NOT EDIT this 360 MARC RECORD

022 a| 2160-7516 y| 2160-7508

035 a| (WaSeSS)ssj0001085593

037 b| Institute of Electrical and Electronics Engineers, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854

040 a| DLC b| eng c| DLC d| OCLCQ d| OCLCA d| OCLCQ d| OCLCA d| DLC d| OCLCO d| OCLCF d| U3W d| OCLCQ d| OCLCL d| WaSeSS

042 a| pcc a| nsdp

050 0 0 a| TA1650

082 1 0 a| 006.4 2| 14 1| <https://id.oclc.org/worldcat/ddc/E3B7T98VPDQQHfYBdbkq8GR9tv>

111 2 a| IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

210 0 a| IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. workshops b| (Online)

222 0 a| IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops b| (Online)

245 1 0 a| IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops h| [electronic resource] : b| [proceedings].

246 1 a| CVPRW

260 a| Piscataway, NJ : b| Institute of Electrical and Electronic Engineers

310 a| Annual

500 a| Proceedings of numerous workshops held in conjunction with CVPR conferences.

538 a| Mode of access: World Wide Web.

588 0 a| Print version record.

588 1 a| 2017 (IEEE Xplore, viewed January 28, 2021).

650 0 a| Optical pattern recognition v| Congresses.

650 0 a| Computer vision v| Congresses.

650 0 a| Pattern recognition systems v| Congresses.

655 0 a| Electronic journals.

710 2 a| Institute of Electrical and Electronics Engineers.

776 0 8 i| Also in print: a| IEEE Computer Society Conference on Computer Vision and Pattern Recognition. t| IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops : [proceedings] x| 2160-7508 w| (DLC) 2011204048 w| (OCoLC)710051718

856 4 0 u| <http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=TC0001085593&T=marc>

910 a| CONSER online

850 a| DLC

949 a| Electronic resource w| ASIS c| 1 i| 48647063-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 8/23/2025 o| .STAFF. 360 MARC

ASK  
A LIBRARIAN

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

## PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#) (814) 865-6368[Libraries Intranet \(Staff Only\)](#)[Accessibility Help](#)[Website Feedback](#)[Policies and Guidelines](#)[Staff Directory](#)

## CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

[Facebook](#)[Twitter](#)[Instagram](#)



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue



# IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops [electronic resource] : [proceedings].

Conference Author

[IEEE Computer Society Conference on Computer Vision and Pattern Recognition](#)

Additional Titles

IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. workshops (Online) and CVPRW

Published

Piscataway, NJ : Institute of Electrical and Electronic Engineers

Additional Creators

[Institute of Electrical and Electronics Engineers](#)

Access Online

[serialsolutions.com](#)

## Subject(s)

[Optical pattern recognition—Congresses](#)[Computer vision—Congresses](#)[Pattern recognition systems—Congresses](#)

## Genre(s)

[Electronic journals](#)

## ISSN

2160-7516

## Publication Frequency

Annual

## Note

Proceedings of numerous workshops held in conjunction with CVPR conferences.

## Technical Details

Mode of access: World Wide Web.

ASK  
A LIBRARIAN[View MARC record](#) | catkey: 48647063

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

## PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#)

(814) 865-6368

[Libraries Intranet \(Staff Only\)](#)[Accessibility Help](#)[Website Feedback](#)[Policies and Guidelines](#)[Staff Directory](#)

## CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

[Facebook](#)[Twitter](#)[Instagram](#)



**PennState**  
University Libraries

RENEW MATERIALS

Give Today

My Account



Services Research About Ask

🕒 Pattee and Paterno: View Hours

## Online Journals, Books and more

☐ All ☒ Journals Only ☐ Books Only ☐ Videos Only

Title begins with

Search

[Show more search options](#) ▼

Showing results 1 through 1 of 1

for the search: "TC0001085593"

Refine Results: [All](#) | [Journals Only](#) | [Books Only](#) | [Videos Only](#)

Limit by: [Peer Reviewed](#) | [Open Access](#)

### IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops

ISSN: 2160-7508

1994 to Present in [IEEE Electronic Library \(IEL\)](#)

Search inside this journal

Search

[Back to Top](#)



**PennState**  
University Libraries



#### SERVICES

Borrow & Renew  
Course Reserves  
Rooms & Spaces  
Technology & Equipment  
Adaptive Technology & Services  
Interlibrary Loan  
Teaching Support for Instructors

#### RESEARCH

Try These First  
Citation Guides  
ScholarSphere  
Collections  
Electronic Theses and Dissertations  
Research Data Services

#### ABOUT

Hours  
Directory  
Libraries and Departments  
Jobs  
Message from the Dean  
Diversity  
News and Events  
Your Donation at Work  
Libraries Calendar

#### ASK A LIBRARIAN

By Chat  
By Phone  
In Person  
By Email  
Schedule a Research Appt.  
FAQs (or How Do I?...)  
Tutorials  
The Search Bar  
Thank a Library Worker

[Libraries Home](#) | [Libraries Intranet \(Employees Only\)](#) | [Accessibility Help](#) | [Website Feedback](#) | [Policies and Guidelines](#) | [Acknowledgement of Land](#) | (814) 865-6368

[Hotline](#) | [Privacy Statements](#) | [Non Discrimination](#) | [Accessibility](#) | [Legal Statements](#)

Copyright ©2025 The Pennsylvania State University.

All rights reserved. Except where otherwise noted, this work is subject to a [Creative Commons Attribution 4.0 license](#), Details and exceptions.

Last Updated on 09-03-2025 © Ex Libris, a ProQuest Company, 2000-2025

# **APPENDIX MOONS01**



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue

## MARC View

LEADER 01231nam a22002652 4500

001 22614831

003 SIRSI

005 20180314130447.0

006 m d

007 cr n

008 160609e20160307nju s||||||| 2|eng|d

020 a| 9781509006427

020 a| 1509006427 (Spiral) c| USD 318.00 Retail Price (Publisher) 9| Active Record

037 b| 00066573

040 a| BIP US d| WaSeSS

110 2 a| IEEE Staff e| Author

245 1 0 a| 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) h| [electronic resource]

260 a| Piscataway : b| IEEE c| March 2016

506 a| License restrictions may limit access.

520 8 a| Annotation b| WACV brings together algorithm developers, software engineers, program managers and others interested in applied computer vision

521 a| Scholarly &amp; Professional b| IEEE

655 0 a| Electronic books.

856 4 0 u| http://SK8ES4MC2L.search.serialssolutions.com/?

sid=sersol&amp;SS\_jc=TC0001947336&amp;title=2016%20IEEE%20Winter%20Conference%20on%20Applications%20of%20Computer%20Vision%20

955 a| Books In Print, (c) 2017 R.R. Bowker LLC

949 a| Electronic resource w| ASIS c| 1 i| 22614831-1001 i| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 3/20/2018 o| .STAFF. 360 MARC

ASK  
A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

### PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#)

(814) 865-6368

[Libraries Intranet \(Staff Only\)](#)[Accessibility Help](#)[Website Feedback](#)[Policies and Guidelines](#)[Staff Directory](#)

### CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

[Facebook](#)[Twitter](#)[Instagram](#)



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue



# 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) [electronic resource]

Corporate Author

[IEEE Staff](#)

Published

Piscataway : IEEE March 2016

Access Online

[serialsolutions.com](#)

## Restrictions on Access

License restrictions may limit access.

## Summary

Annotation WACV brings together algorithm developers, software engineers, program managers and others interested in applied computer vision

## Genre(s)

[Electronic books](#)

## ISBN

9781509006427

1509006427 (Spiral)

## Audience Notes

Scholarly &amp; Professional IEEE

ASK  
A LIBRARIAN[View MARC record](#) | catkey: 22614831

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

## PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#)

(814) 865-6368

[Libraries Intranet \(Staff Only\)](#)[Accessibility Help](#)[Website Feedback](#)[Policies and Guidelines](#)[Staff Directory](#)

## CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

[Facebook](#)[Twitter](#)[Instagram](#)

# **APPENDIX MOONS02**

# Energy-Efficient ConvNets Through Approximate Computing

Bert Moons<sup>†\*</sup> Bert De Brabandere<sup>°\*</sup> Luc Van Gool<sup>°</sup> Marian Verhelst<sup>†</sup>  
<sup>†</sup>ESAT-MICAS, <sup>°</sup>ESAT-VISICS, Dept. of Electrical Engineering, KU Leuven  
bert.moons@esat.kuleuven.be, bert.debrabandere@esat.kuleuven.be

## Abstract

*Recently convolutional neural networks (ConvNets) have come up as state-of-the-art classification and detection algorithms, achieving near-human performance in visual detection. However, ConvNet algorithms are typically very computation and memory intensive. In order to be able to embed ConvNet-based classification into wearable platforms and embedded systems such as smartphones or ubiquitous electronics for the internet-of-things, their energy consumption should be reduced drastically. This paper proposes methods based on approximate computing to reduce energy consumption in state-of-the-art ConvNet accelerators. By combining techniques both at the system- and circuit level, we can gain energy in the systems arithmetic: up to  $30\times$  without losing classification accuracy and more than  $100\times$  at 99% classification accuracy, compared to the commonly used 16-bit fixed point number format.*

## 1. Introduction

Recently neural networks have made an impressive comeback in the field of machine learning. Convolutional neural networks or ConvNets are consistently pushing the state-of-the-art in areas like computer vision and speech processing. One of the reasons for this revival is the increasing availability of computing power. Multicore CPU's, GPU's, and even clusters of GPU's are no longer prohibitively expensive and make it possible to train and evaluate larger networks.

Unfortunately, the increase in computing power is inevitably accompanied by an increase in energy consumption. While high energy consumption is no big concern during the network's training phase - which typically takes place on a computer cluster - it poses a problem when the network needs to be evaluated on mobile hardware like smartphones, smart glasses and other wearable devices.

In this paper, we investigate how energy consumption can be reduced through principles of approximate computing.

Our analysis shows that it is possible to quantize existing networks and drastically reduce the number of bits that encode the weights and inputs of each layer, with only a minimal loss in accuracy and without the need to retrain the networks. With an appropriate hardware architecture this knowledge can be used to reduce energy consumption in many neural network applications and even dynamically scale energy consumption depending on the current application.

Our contribution is threefold:

1. We show how computational precision can be scaled in several ConvNet architectures for image classification through quantization of the layer's inputs and weights. The possible quantization varies per architecture, per application and even per layer within a single ConvNet.
2. We show how this approximate computing / precision scaling can lead to reduced energy consumption in ConvNet accelerators. This is achieved through a combination of algorithmical and circuit-level techniques:
  - Applying precision scaling on ConvNet-accelerator circuits.
  - The number of zero-valued parameter and input-values of ConvNet layers increases through precision scaling. Their computations can be skipped.
3. Based on this analysis of algorithmic accuracy and energy savings, we demonstrate achievable energy-accuracy curves for three popular ConvNet architectures for image classification.

## 2. Related work

While ConvNets have a long history in the field of computer vision, only recently they have become the go-to technique for classification and detection problems. This newfound popularity is a result of their impressive performance on a number of benchmarks, reaching super-human accuracy on tasks like handwritten character recognition [13] and near-human performance on large scale classification

\*Equally contributing authors.

challenges like ILSVRC [4]. Record-breaking ConvNet architectures are being developed in academia [10] [16] and industry [17] [21].

A ConvNet has a hierarchical structure consisting of a number of layers. Most architectures contain multiple blocks of convolutional layers, rectified linear units (ReLU) and pooling layers, followed by a number of fully connected layers. The network is trained with the backpropagation algorithm, which iteratively updates the weights in the convolutional and fully connected layers in order to minimize a certain loss function.

The success of these algorithms has led to the interest of the embedded vision community. A number of ConvNet hardware optimizations [2] [15], coprocessors [6] and accelerators such as DianNao [1] have been proposed. Although all of these works meant a significant step forward, the realization of a high-performance, energy-efficient and real-time ConvNet architecture, which could work autonomously from the cloud is not yet complete. The optimized architectures named above, can however be optimized further by exploiting ConvNet's inherent error-resilience. The effect of quantization errors on regular neural networks have been studied in [8] and [5].

Error resilient applications can be made more energy efficient through approximate computing. This is a collection of hardware-level techniques enabling energy savings at the expense of reduced computational accuracy. This can be achieved in various ways, leading to either a static (fixed after design-time) or a dynamic (adaptable after design-time) trade-off. Static examples in literature make use of approximate digital building blocks for arithmetic functions [11], or use approximate logic synthesis methods [20]. Dynamic approximate computing mainly uses the principle of precision scaling, in which the number of bits used in computations is varied at run-time [18] [3] [14]. In this work, we use precision scaling to achieve energy gains in ConvNet acceleration. Since precision scaling can be applied dynamically, the energy consumption of the targeted accelerator can be adapted to the varying precision requirements of the used ConvNet and even of the ConvNet layer.

Approximate computing through precision scaling lowers the active power consumption of a digital circuit. The total power consumption consists of a dynamic and a leakage component. It can be summarized as  $P = \alpha C f V^2 + P_{leakage}$  where  $\alpha$  is the circuit switching activity,  $f$  is the clock frequency,  $C$  is the total switching capacitance and  $V$  is the circuit's supply voltage. By scaling precision (i.e., dynamically scaling the number of bits that encode the network's weights and inputs), the switching activity  $\alpha$  can be substantially reduced. Figure 1 illustrates the achievable energy-accuracy trade-off in a typical digital multiplier. Note that a striking  $12\times$  energy gain can be achieved at an output error of 1% root-mean-square-error (RMSE).

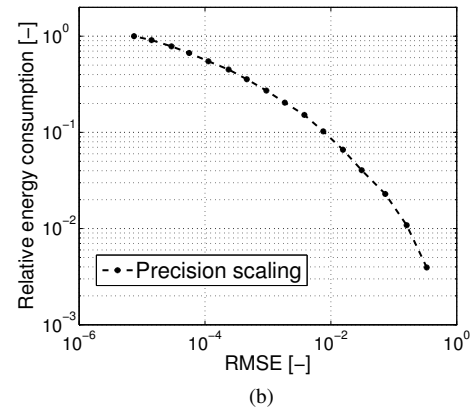
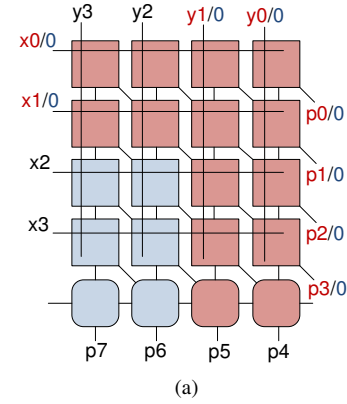


Figure 1: Energy-accuracy trade-off in digital multipliers through precision scaling (a) By scaling the used number of bits from 4 to 2 the number of active blocks drops from 20 to 6. (b) Energy consumption versus root-mean-square-error under precision scaling. Precision scaling can gain  $12\times$  in energy at 1% RMSE.

A first attempt in combining neural networks with approximate computing has been done in [19]. Their work only looks into regular neural networks and requires retraining of the network. In this work we focus on convolutional neural network architectures, which perform much better in computer vision applications. Furthermore, we do not require network retraining.

### 3. Analysis and experiments

We examine the performance- and energy-related effects of quantization on three different network architectures for image classification. Because it is impossible to cover the entire spectrum of architectures and applications, we choose three popular networks as representatives of small-, medium-, and large-scale architectures that differ significantly in their number of parameters:

- LeNet-5 on MNIST [12]: LeNet-5 is a small network

with two convolutional and two fully connected layers. It is designed to classify handwritten digits (28x28 images) in the MNIST dataset where it achieves an accuracy of 99.0%.

- **CifarQuick on Cifar-10 [9]:** CifarQuick is a medium-sized architecture with three convolutional and two fully connected layers. It has more filters per layer than LeNet-5 and operates on color images instead of grayscale images. It classifies images of the Cifar-10 dataset, which consists of small 32x32 color images divided into ten classes. It reaches an accuracy of 75.3%.
- **AlexNet on ImageNet [10]:** AlexNet is a large network with five convolutional and three fully connected layers. It reaches a top-5 accuracy of 80.0% on ILSVRC2012, a large scale classification challenge with 1000 categories.

For our experiments we customized the open-source deep learning framework Caffe [7] to be able to simulate quantization of the network's weights and inputs. All experiments are run on the validation sets of the discussed benchmarks. We always report the relative accuracy, i.e. the ratio of the accuracy after quantization and the accuracy of the original network.

Note that our techniques can be used to reduce the energy consumption during evaluation of the network, but not necessarily during the training phase. When training the network, the weight quantization would cause the backpropagation algorithm to quickly get stuck in local optima.

### 3.1. Influence of quantization on classification accuracy

Typically, ConvNets run on high precision machines using 32-bit floating point number representations. Dedicated embedded platforms use 16-bit fixed point hardware for ConvNet computations. However, such high precision is not always necessary in. The energy spent in high precision computations, does not lead to more accurate classification by the algorithm. To reduce the energy consumption of the ConvNet's computations, our main strategy is to quantize its weights and the inputs to its layers. Such quantization leads to a network that is only an approximation of the original network. Our goal is to find out the influence of quantization on a network's accuracy and whether the effects differ significantly across network architectures. We are interested in how far we can push the quantization of a network without sacrificing too much classification accuracy.

Before quantizing the weights and inputs, it is important to rescale them properly according to the distribution of their values. If there is a mismatch between the interval in which these values lie and the interval over which we quantize, the accuracy will drop even at very low quantization settings. For this reason, we rescale all layer inputs and

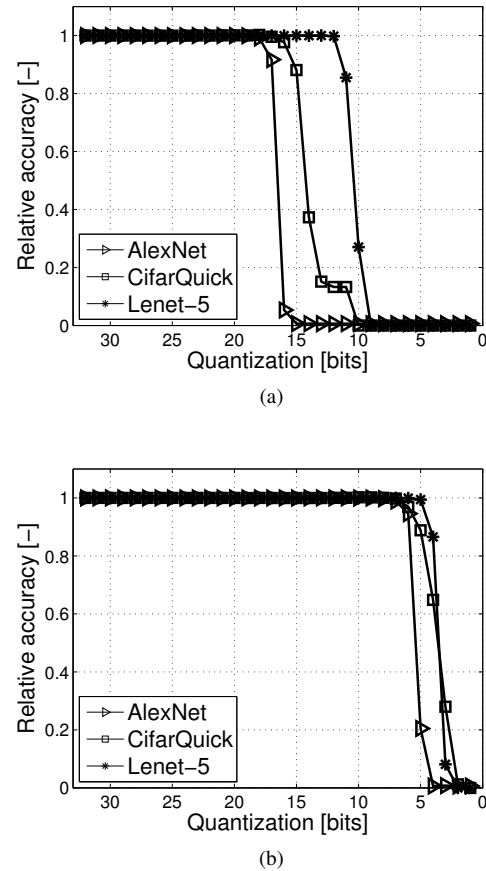


Figure 2: Relative accuracy of three architectures as a function of the number of quantization bits. Layer inputs and weights are first rescaled so they can be quantized more effectively. We compare two rescaling strategies: (a) Uniform rescaling: all inputs and weights are scaled with the same value. (b) Per-layer rescaling: inputs and weights are scaled on a per-layer basis.

weights in the network with a single value that corresponds to the maximal input or weight value (rounded to the next factor of two) observed during a complete run over the data. This ensures that the limits of the quantization interval correspond to the limits of the data, and no bits are wasted. Mathematically, this scaling is the same as multiplying the results with a factor of 2, or shifting data in a fixed point number representation. This operation has a limited hardware energy footprint and only needs to be performed at convolutional outputs.

#### 3.1.1 Uniform quantization

As a first experiment we choose a single quantization setting for all layers in the network, and call this *uniform quantization*. Later, we also introduce *per-layer quantization*, where

each layer is quantized separately.

Figure 2a shows the relative accuracy of our three networks as a function of the number of quantization bits. As can be seen in the figure, the relative accuracy stays equal to one for all three networks up to quantization with 19 bits, meaning that the quantized network reaches the exact same accuracy on its dataset as the original network. At a quantization with 18-bit however, the accuracy for AlexNet starts dropping quickly, rendering it useless for any practical application. At 11-bit the same effect can be seen for the smaller LeNet-5.

We can do better than this by applying the scaling in a more fine-grained way; we choose a different scaling factor for each layer. This is clarified by a simple example: AlexNet’s first and sixth layer weight statistics are shown in Figure 3. All weights of layer 1 are within the  $[-0.5, 0.5]$  interval, while the weights of layer 6 are within the  $[-0.0625, 0.0625]$  interval. By allowing to quantize the weights in layer 6 over this smaller interval instead of over  $[-0.5, 0.5]$  we recover 3 bits that would otherwise be wasted. A similar reasoning holds for the layer’s inputs.

The effect of the per-layer scaling is significant, as can be seen in Figure 2b. Compared to the previous, we can now quantize much more aggressively without sacrificing accuracy. Each network performs at its original accuracy up to quantization with no more than 8 bits. After that, the accuracy again drops quickly. The reason for this improvement is the fact that input and weight statistics differ greatly among layers. Per-layer rescaling allows to set the optimal quantization interval in each layer.

### 3.1.2 Per-layer quantization

Just as we do per-layer rescaling, we can also do per-layer quantization: instead of quantizing all weights and inputs of the network with the same number of bits, we choose a different setting in each layer. The idea is again to find an optimal setting for each layer by exploiting the variations of the layer-specific input and weight distributions. Another effect of the per-layer quantization is that we can set the operating point (i.e. the desired minimal relative accuracy) more precisely and thus control the energy-accuracy trade-off tightly, as discussed later in section 3.3. This is not possible with uniform quantization: e.g. between quantization with 5 and 4 bits, the relative accuracy of LeNet-5 drops from 99.4% immediately to an unusable 86.6%.

In order to find a good quantization setting for each layer, we do a greedy search over the parameters: starting at the first layer, its input is quantized until the accuracy drops to the target accuracy. Next, the quantization of the input is kept fixed while the quantization of its weights is maximized in the same way. The same process is applied in the next layers until the last one. If a value goes out of range,

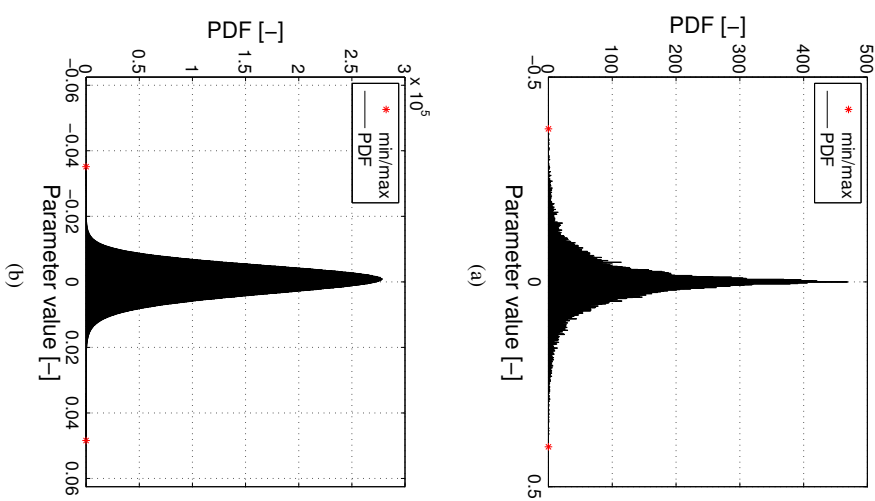


Figure 3: AlexNet’s first and sixth layers weight statistics. Layer 1 should be projected on the  $[-0.5, 0.5]$  interval and layer 6 on the  $[-0.0625, 0.0625]$  interval to minimize the needed number of bits. If layer 6 would be projected on the  $[-1, 1]$  interval, 4 extra bits would be needed.

we clip it to the respective minimum or maximum representable value with the chosen quantization. The full sweep can take many hours, depending on the network size, dataset size and number of tested settings. This poses no real problem, as the procedure has to be performed only once and can be done off-line.

The amount of bits that can be saved with per-layer quantization at a target accuracy of 99%, compared to uniform quantization at 100% relative accuracy, is visualized for each reference network in Figure 4. The results are ad-hoc, but there is a general trend of needing less bits in the lower layers of the network than in the higher layers. This is partly a result of the forward parameter sweep, but we hypothesize that the difference in input and weight statistics between lower and higher layers also plays a significant role.

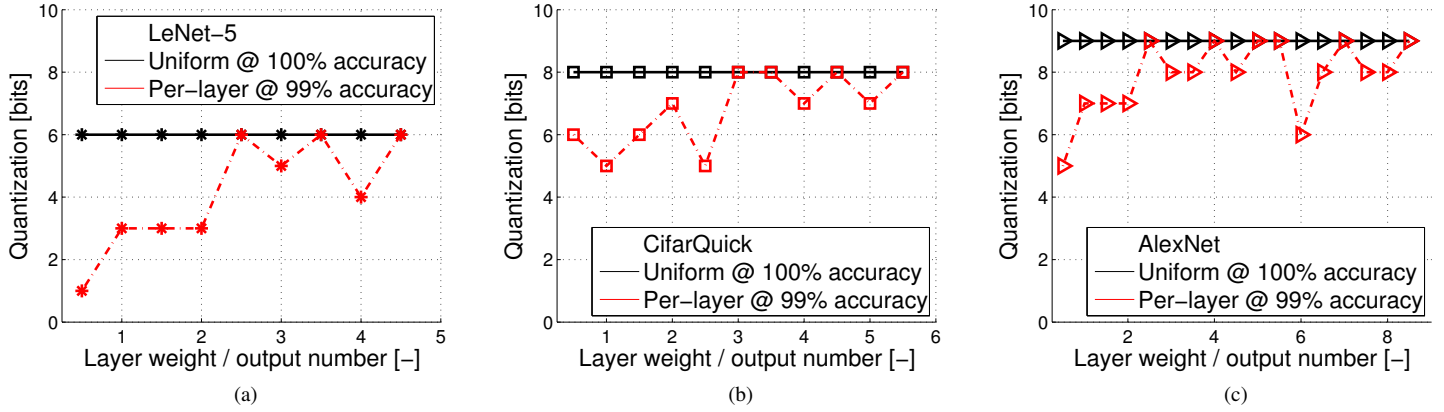


Figure 4: Comparison of uniform and optimal per-layer quantization. Data points at integer values show quantization of the layer weights. Other data points show quantization of the layer inputs. Maximum achievable quantizations: (a) LeNet-5: 1-6 bit. (b) CifarQuick: 5-8 bit. (c) AlexNet: 5-9 bit.

### 3.2. Proposed energy reduction through precision scaling and computation skipping

This section indicates how this increased quantization can lead to energy savings in real hardware architectures. To quantify the possible energy gains of approximate computing in ConvNet acceleration, we model the energy consumption of the necessary convolutional arithmetic (multiply and add) for a full ConvNet algorithm. This analysis does not incorporate the energy overheads of control, I/O, data- and program-memory interfacing and the clock network, as these are highly architecture dependent. It can thus be considered as the maximum potential energy savings through approximate computing. Furthermore, the number of zero-valued weights and inputs of typical ConvNets increases at higher quantization. This can be exploited algorithmically by skipping unnecessary computations with zero-valued inputs. This will lead to large additional energy savings.

#### 3.2.1 Precision scaling

Precision scaling is an approximate computing technique allowing major energy savings in digital circuits. Lower precision computing, i.e. computing using less bits, reduces the switching activity. Figure 1 and [14] show how this concept can be applied to a common digital multiplier.

Based on simulations in a 40nm technology, an energy model for common digital building blocks is proposed in Table 1. The reduced switching activity, and hence energy consumption, solely depends on the circuit architecture and can reduce quadratically (multiplier), or linearly (adders, register files, wiring). Using the energy-modelling from Ta-

Table 1: First order estimation of power consumption in digital basic building blocks under accuracy scaling, based on simulations in a 40nm CMOS technology. The modelled blocks are multipliers ( $\times$ ), adders ( $+$ ), registers, wires and SRAM memory. In the power equation ( $P$ ), only the switching activity  $\alpha$  decreases as a function of the used number of bits  $n$ .

| $P = \alpha C f V_{dd}^2$ | $\times$ | $+$    | Reg.   | Wire   | SRAM         |
|---------------------------|----------|--------|--------|--------|--------------|
| $\alpha$                  | $O(n^2)$ | $O(n)$ | $O(n)$ | $O(n)$ | $O(n_{max})$ |

ble 1, we can estimate the impact of increased quantization arithmetic on the energy consumption of a full ConvNet-algorithms.

In order to apply precision scaling efficiently in such a system, architectural changes have to be made. First, the positioning of the sign-bit in the fixed number representation is crucial. The MSB-bit should always be placed at the MSB position, otherwise toggling sign bits will lead to high switching activity. Second, The number of bits in a fixed point implementation tends to expand due to numerical operations. In order to make sure the number of bits remains limited throughout different stages of the algorithm, repetitive rounding or truncation is needed after every stage. Third, data- and parameter rescaling should be implemented between the convolutional layers using arithmetic shifting.

#### 3.2.2 Computation skipping

An interesting feature of many modern convolutional neural networks is the appearance of Rectified Linear Unit

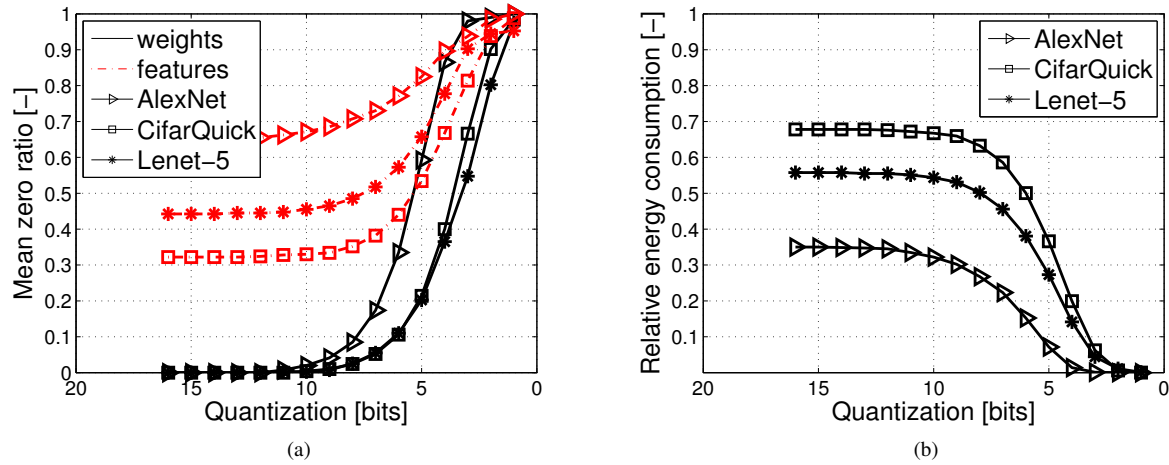


Figure 5: (a) Impact of precision scaling on the average number of zeroes in the weights and input values of all three examples. (b) Relative energy reduction compared to the case without computation skipping.

layers (ReLU layers). These put all negative inputs to zero and pass on positive values unchanged, as in  $Output = \max(0, Input)$ . Since many layers in ConvNet-classification algorithms only output positive values when certain features are present, a large amount of ReLU-outputs will be zero and do not have to be used for further computations. The ReLU layers thus allow for additional energy reductions by not computing unnecessary computations through computation skipping.

Figure 5 shows the impact of precision scaling on the average number of zeroes in our ConvNet examples. For all architectures, the ratio of zeroed values lies between 50 – 90% of the total input values, depending on the used quantization setting. The number of zeroed values is significantly larger under precision scaling, from an average of 45% on all LeNet-5 outputs at 16-bit to 65% at 5-bit.

It is very difficult to achieve such computation skipping in a software solution, since checking for zero-valued inputs in a scalar core consumes time. Using techniques for sparse matrix multiplication is not possible either, because the used convolutional kernels are typically very small on the order of  $11 \times 11$ ,  $5 \times 5$  or  $3 \times 3$ . However, computation skipping can be achieved in a hardware accelerator with dedicated hardware support. Flags can indicate if upcoming data is zero and prevent circuitry from switching if this is the case. As ConvNet-weights are fixed, these flags can be computed beforehand.

### 3.3. Achievable energy-accuracy trade-off in ConvNets

By combining both precision scaling and algorithmical computation skipping, major energy savings can be achieved in ConvNets. Figure 6a shows the effect of uni-

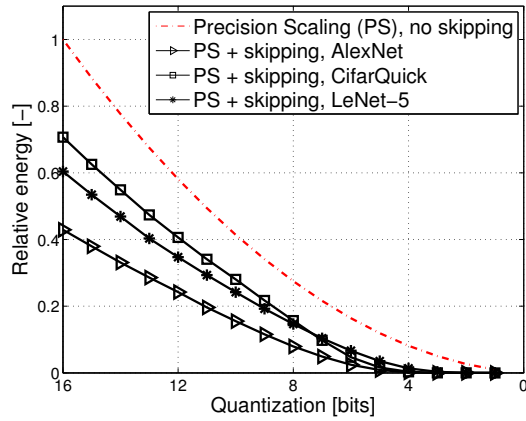
form quantization on the energy consumption of our benchmark algorithms, if both precision scaling and computation skipping are combined. Note how the energy reduces 5 – 20× for an 8-bit implementation.

Figure 6b shows curves in the energy-accuracy space for our ConvNet benchmarks. It shows the trade-off for the 95 – 100% classification accuracy window. All curves are given for both uniform and per-layer quantization, and are compared to the typically used uniform 16-bit number representation, without adequate precision scaling and computation skipping. If per-layer quantization is used, the trade-off is less steep than in the uniform case. This means more energy can be gained, while losing less classification accuracy in the process. All discussed networks gain at least an additional factor of 2 in energy consumption if a reduced classification accuracy of 99% is allowed.

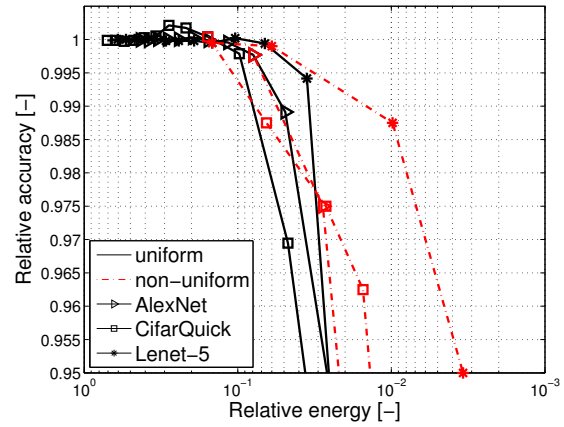
Figure 7 gives an overview of the effect of each improvement on the relative energy consumption. Per-layer rescaling and computation skipping lead to the largest energy gains. Per-layer quantization leads to additional major energy gains if a reduced accuracy of 99% can be allowed.

## 4. Conclusion

ConvNet algorithms are typically very computation and memory intensive. In order to be able to embed ConvNet-based classification into wearable platforms, their energy consumption should be reduced significantly. In this work, we show how the energy consumption of dedicated ConvNet accelerators can be drastically reduced by applying approximate computing. ConvNets for image classification are fault-tolerant and allow computation using 4-10 bits instead of the previously used 16- or 32-bit formats. By us-

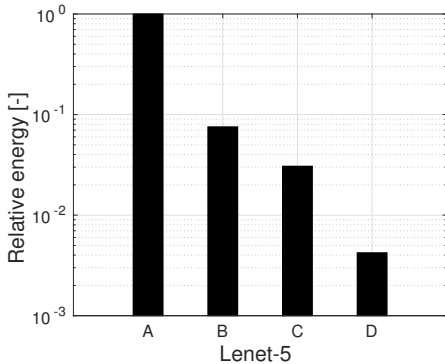


(a)

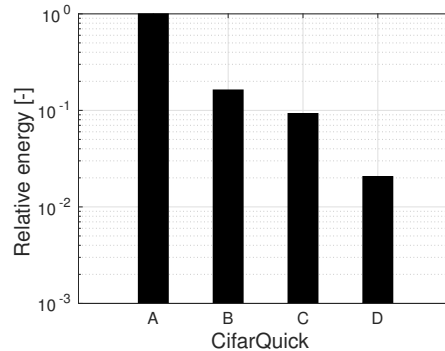


(b)

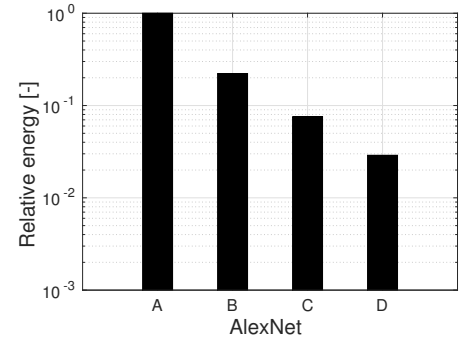
Figure 6: (a) Relative energy consumption for uniform quantization. A striking 5 – 20 $\times$  lower energy consumption at 8-bit, compared to the commonly used 16-bit used in previous literature. (b) Energy-accuracy plot for per-layer quantization and scaling and with computation skipping. All considered ConvNet-networks gain at least an additional factor of 2 in energy consumption if 99% accuracy is allowed.



(a)



(b)



(c)

Figure 7: Energy comparison between our three benchmarks in different cases. (A) 16-bit (B) Per-layer scaling, uniform quantization at 100% accuracy (C) Same as B, with computation skipping (D) Per-layer scaling, per-layer quantization and computation skipping at 99% accuracy.

ing requantized number representations, the energy consumption in a ConvNet accelerator can be reduced in two complementary ways. First, energy consumption can be lower through precision scaling. Second, more quantization further increases the number of zeroed parameter and input values in the ConvNet. Since zero-valued numbers do not contribute to the algorithm's output, their computations can be skipped. The combination of these two techniques can lead to significant energy savings in several ConvNets. We show energy reductions of up to 30 $\times$  compared to the commonly used 16-bit fixed point implementations, with-

out sacrificing algorithm performance. If the classification accuracy can be reduced to 99%, additional energy savings can be achieved on the same network architecture, through per-layer optimization. In this case an additional 7.5 $\times$  reduction can be achieved.

## References

- [1] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In *Proceedings of the 19th international conference on Architectural support*

- for programming languages and operating systems, pages 269–284. ACM, 2014. 2
- [2] F. Conti and L. Benini. A ultra-low-energy convolution engine for fast brain-inspired vision in multicore clusters. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, pages 683–688. EDA Consortium, 2015. 2
- [3] M. de la Guia Solaz, W. Han, and R. Conway. A flexible low power dsp with a programmable truncated multiplier. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 59(11):2555–2568, 2012. 2
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pages 248–255, June 2009. 2
- [5] G. Dundar and K. Rose. The effects of quantization on multilayer neural networks. *IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council*, 6(6):1446–1451, 1994. 2
- [6] C. Farabet, B. Martini, B. Corda, P. Akselrod, E. Culurciello, and Y. LeCun. Neuflow: A runtime reconfigurable dataflow processor for vision. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 109–116. IEEE, 2011. 2
- [7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093 [cs]*, June 2014. arXiv: 1408.5093. 3
- [8] M. Jiang and G. Gielen. The effects of quantization on multi-layer feedforward neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(04):637–661, June 2003. 2
- [9] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. 3
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems* 25, pages 1097–1105. Curran Associates, Inc., 2012. 2, 3
- [11] P. Kulkarni, P. Gupta, and M. Ercegovac. Trading accuracy for power with an underdesigned multiplier architecture. In *VLSI Design (VLSI Design), 2011 24th International Conference on*, pages 346–351. IEEE, 2011. 2
- [12] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. Citeseer, 1990. 2
- [13] Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998. 1
- [14] B. Moons and M. Verhelst. Dvas: Dynamic voltage accuracy scaling for increased energy-efficiency in approximate computing. *International Symposium on Low Power Electronics and Design (ISLPED)*, 2015. 2, 5
- [15] M. Peemen, A. A. Setio, B. Mesman, and H. Corporaal. Memory-centric accelerator design for convolutional neural networks. In *Computer Design (ICCD), 2013 IEEE 31st International Conference on*, pages 13–19. IEEE, 2013. 2
- [16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2
- [17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*, Sept. 2014. arXiv: 1409.4842. 2
- [18] S. Venkataramani, V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan. Quality programmable vector processors for approximate computing. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 1–12. ACM, 2013. 2
- [19] S. Venkataramani, A. Ranjan, K. Roy, and A. Raghunathan. Axnn: energy-efficient neuromorphic systems using approximate computing. In *Proceedings of the 2014 international symposium on Low power electronics and design*, pages 27–32. ACM, 2014. 2
- [20] S. Venkataramani, A. Sabne, V. Kozhikkottu, K. Roy, and A. Raghunathan. Salsa: systematic logic synthesis of approximate circuits. In *Proceedings of the 49th Annual Design Automation Conference*, pages 796–801. ACM, 2012. 2
- [21] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. Deep Image: Scaling up Image Recognition. *arXiv:1501.02876 [cs]*, Jan. 2015. arXiv: 1501.02876. 2

# **APPENDIX KAUL01**



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue

## MARC View

LEADER 01349nam a22002652 4500

001 20057365

003 SIRSI

005 20170414130234.0

006 m d

007 cr n

008 160217e20160131nju s||||| 2|eng|d

020 a| 9781467394680

020 a| 1467394688 (Spiral) c| USD 365.00 Retail Price (Publisher) 9| Active Record

037 b| 00066573

040 a| BIP US d| WaSeSS

110 2 a| IEEE Staff e| Author

245 1 0 a| 2016 IEEE International Solid State Circuits Conference (ISSCC) h| [electronic resource]

260 a| Piscataway : b| IEEE c| Jan. 2016

506 a| License restrictions may limit access.

520 8 a| Annotation b| foremost forum for presentation of advances in solid state circuits and systems on a chip The Conference offers a unique opportunity for engineers working at the cutting edge of IC design and use to maintain technical currency, and to network with leading experts

521 a| Scholarly &amp; Professional b| IEEE

655 0 a| Electronic books.

856 4 0 u| [http://SK8ES4MC2L.search.serialssolutions.com/?](http://SK8ES4MC2L.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001773180&title=2016%20IEEE%20International%20Solid%20State%20Circuits%20Conference%20%28ISSCC%29)[sid=sersol&SS\\_jc=TC0001773180&title=2016%20IEEE%20International%20Solid%20State%20Circuits%20Conference%20%28ISSCC%29](http://SK8ES4MC2L.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001773180&title=2016%20IEEE%20International%20Solid%20State%20Circuits%20Conference%20%28ISSCC%29)

955 a| Books In Print, (c) 2017 R.R. Bowker LLC

949 a| Electronic resource w| ASIS c| 1 i| 20057365-1001 I| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 4/18/2017 o| .STAFF. 360 MARC

**ASK**  
A LIBRARIAN 

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

### PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#) (814) 865-6368  
[Libraries Intranet \(Staff Only\)](#)  
[Accessibility Help](#)  
[Website Feedback](#)  
[Policies and Guidelines](#)  
[Staff Directory](#)

### CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook  
 Twitter  
 Instagram



Keyword



Search...

Search

[Advanced search](#) [Start Over](#)

Share

Bookmark

Report an Issue



## 2016 IEEE International Solid State Circuits Conference (ISSCC) [electronic resource]

Corporate Author

[IEEE Staff](#)

Published

Piscataway : IEEE Jan. 2016

### Access Online

[serialsolutions.com](#)

### Restrictions on Access

License restrictions may limit access.

### Summary

Annotation foremost forum for presentation of advances in solid state circuits and systems on a chip The Conference offers a unique opportunity for engineers working at the cutting edge of IC design and use to maintain technical currency, and to network with leading experts

### Genre(s)

[Electronic books](#)

### ISBN

9781467394680

1467394688 (Spiral)

### Audience Notes

Scholarly &amp; Professional IEEE

ASK  
A LIBRARIAN[View MARC record](#) | catkey: 20057365

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

[Legal Statements](#) | [PSU Hotlines](#)

### PENN STATE UNIVERSITY LIBRARIES

[Libraries Home](#)

(814) 865-6368

[Libraries Intranet \(Staff Only\)](#)[Accessibility Help](#)[Website Feedback](#)[Policies and Guidelines](#)[Staff Directory](#)

### CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

[Facebook](#)[Twitter](#)[Instagram](#)

# **APPENDIX KAUL02**

#### 14.4 A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS

Himanshu Kaul, Mark A. Anders, Sanu K. Mathew, Gregory Chen, Sudhir K. Satpathy, Steven K. Hsu, Amit Agarwal, Ram K. Krishnamurthy

Intel, Hillsboro, OR

Energy-efficient k-nearest-neighbor (kNN) computations are key building blocks for computer vision, classification, and machine-learning workloads [1-3]. Determining distances to high-dimensional vectors within a large vector database results in high compute cost. Adaptive precision improves energy efficiency by eliminating a majority of vectors without costly full-precision computation, with as-needed precision refinement to guarantee kNN accuracy of closely matched vectors. A special-purpose on-die kNN accelerator with 128-dimensions by 128 parallel reference vectors, targeted across mobile SoCs to multi-core microprocessors, and reconfigurable for either Manhattan or Euclidean distance, is fabricated in 14nm tri-gate CMOS [6]. Partial distance compute circuits, 2b window-based sort, MSB-to-LSB-based selective distance refinement, robust ultra-low voltage circuits, and state tracking control to selectively resume next-nearest candidates enable nominal energy efficiency of 3.37nJ/query vector or 9.7TOPS/W (measured for 21.5M vectors/s, 16 cycles/vector at 750mV, 25°C) with a dense layout occupying 0.333mm<sup>2</sup> (Fig. 14.4.7) while achieving: i) scalable performance up to 26.4M vectors/s, 114mW measured at 850mV, ii) 2-cycle latency and 43pJ energy to find each subsequent nearest neighbor, iii) up to 5.2× higher throughput while maintaining full-precision kNN accuracy, iv) 16× search-space reduction for next-nearest neighbor, v) ultra-low voltage operation measured at 360mV, 1.1M vectors/s, 1.44mW, and vi) peak energy efficiency of 1.23nJ/vector at 390mV (near-threshold), 25°C.

The kNN accelerator finds the nearest-neighbor indices using in-memory distance-compute circuits, a distributed sort network and efficient incremental search (Fig. 14.4.1). Parallel Manhattan and Euclidean distance computations to the 128 reference vectors (128-D per vector) start with low-accuracy from the MSB using partial distance compute circuits. The sort network finds the minimum within current accuracy bounds, eliminating vectors outside these bounds for up to 6× compute reduction. The partial distance compute circuits refine the remaining distances, with a successive 2b of the accumulated distance available for further sorting in each iteration. Search-space control circuits coordinate these partial distance iterations, vector eliminations and 2b sorting windows until the nearest neighbor is found. Since the last eliminated group of vectors will contain the next nearest neighbor, tracking these groups enables an efficient incremental search within only the last group. Early vector elimination improves throughput up to 5.2×, and narrow datapath width reduces accelerator area by 46%, while maintaining full-precision kNN accuracy.

Complete Euclidean and Manhattan distances from dimension-wise 8b absolute differences require accumulation over 6 and 4 partial-compute iterations, respectively (Fig. 14.4.2a). Iterative refinement of accumulated distance is interleaved as needed with successive sort operations on 2b windows from MSB to LSB. Area-optimized absolute difference uses multi-bit latches for reference vector storage and ripple-carry static mirror adders (Fig. 14.4.2b). The MSB carry-out connects to the LSB carry-in for 2's complement absolute value increment, replacing a separate incrementer for 30% lower area. Use of partial-squaring operations reduce each 8b multiplier to a pair of 2b multipliers, with the shuffle circuit selecting the appropriate inputs from the absolute difference. Narrow partial distance bit-widths also reduce compressor tree area and energy. The 4b products from each dimension are added separately to relocate the left-shift and zero-select circuits after the compressor tree, reducing them from one per dimension to one per vector. The shared partial-squaring and compressor-tree circuits are multiplexed between every pair of vectors using the *vector select* signal for 28% lower area. The previously accumulated distance is shifted left 2b in every compute iteration, enabling replacement of an 11b right shift in the critical path for incremental distance alignment with only a 1b right shift (Fig. 14.4.3). This also decreases the radix-2 carry-merge adder width to 12b for incremental distance summation, with a non-critical 10b incrementer for the upper bits. These optimizations reduce accumulator delay by 37% and area by 48%. Each distance refinement can affect previously sorted 2b windows by at most 1, which must be included in each partial-sort iteration. Therefore, the 12:1 multiplexers forward a

carry-out or prior difference (Y) along with the correct accumulator 2b window to the sort network. The *precise* signals indicate whether all other *psums* exceed the minimum by more than 1, signifying that the same minimum has been found as with full-precision calculations. Decreasing comparator width from 23b to 3b simultaneously reduces sort network area by 3.6× and latency by 2×. Overall energy to find the nearest neighbor reduces up to 3× compared to a full-precision design.

Search-space control enables tracking: i) the vectors that could be possible winners until only one remains, ii) the current 2b window indicated by a global pointer, and iii) 2b windows with eliminated vectors for the next nearest neighbor search. The control-circuit location is divided among vector-compute blocks and a centralized global control (Fig. 14.4.4). The *compare psum* circuit uses the global minimum to determine vector elimination and nearest neighbor selection. Early removal of vectors while moving MSB to LSB for nearest-neighbor search results in 47% compute reduction and a 2b distance sort reduction of 3.8×. The global binary mask stores a one when vectors are eliminated as the global pointer advances MSB to LSB. The global pointer rolls back to the closest 2b window in this binary mask since the vectors removed in this window contain the next-nearest neighbor. Each vector stores the pointer position at elimination, and during rollback the *compare pointer* circuits re-enable only the closest vectors. This search for the next nearest neighbor within a selective subset of the remaining vectors results in 16× search-space reduction.

The adaptive precision kNN accelerator finds the nearest neighbor out of 128×128-D randomly generated vectors for 21.5M query vectors/s at 16 cycles/vector with Manhattan distance measured at 338MHz, nominal 750mV, 25°C (Fig. 14.4.5). Corresponding total power consumption is 73mW (1.8mW active leakage component), resulting in energy-per-query vector of 3.37nJ or 9.7TOPS/W (10P = 8b add/multiply/compare). As k scales from 1 to 10, mean increase in latency and energy per k is 2 cycles and 43pJ. Nominal Euclidean operation is 20.3M vectors/s at 17 cycles/vector, consuming 78mW or 3.84nJ/vector. Mean increase in latency and energy per k is 4 cycles and 87pJ. Peak Manhattan (Euclidean) performance scales up to 26.4M (24.8M) vectors/s with power consumption of 114mW (123mW) at 850mV (Fig. 14.4.6). Datapath circuits optimized for reliable ultra-low-voltage operation [4] enable robust functionality measured down to 360mV, with k=1 throughput scaling to 1.1M vectors/s, 1.44mW for Manhattan and 1.04M vectors/s, 1.63mW for Euclidean computations. Energy decreases with supply voltage for a peak efficiency of 1.23nJ per query vector or 26.5TOPS/W with Manhattan (k=1) mode at 390mV (near-threshold), representing a 2.73× energy efficiency improvement over nominal 750mV operation. Energy-efficiency measurements also improve with algorithm-dependent SURF-128 [5] scale- and rotation-invariant feature vector matching for object recognition. These feature vectors achieve an additional 15% (Manhattan) and 20% (Euclidean) energy savings at 750mV, resulting in 2.86nJ/vector and 3.07nJ/vector, respectively.

##### Acknowledgments:

The authors thank M. Haycock, M. Mayberry, H. Jiang, S. Borkar, V. De, M. Kounavis and F. Ghenassia for encouragement and discussions. This research was, in part, funded by the U.S. Government. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

##### References:

- [1] Y. Kim et al., "A 0.5V 54μW Ultra-Low-Power Recognition Processor with 93.5% Accuracy Geometric Vocabulary Tree and 47.5% Database Compression," *ISSCC Dig. Tech. Papers*, pp. 330-331, 2015.
- [2] I. Hong et al., "A 125,582 vector/s Throughput and 95.1% Accuracy ANN Searching Processor with Neuro-Fuzzy Vision Cache for Real-time Object Recognition," *IEEE Symp. on VLSI Circuits*, pp. 184-185, 2013.
- [3] T.-W. Chen et al., "A Multimedia Semantic Analysis SoC (SASoC) with Machine-Learning Engine," *ISSCC Dig. Tech. Papers*, pp. 338-339, 2010.
- [4] S. Hsu et al., "A 280mV-to-1.1V 256b Reconfigurable SIMD Vector Permutation Engine with 2-Dimensional Shuffle in 22nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 178-179, 2012.
- [5] H. Bay et al., "SURF: Speeded Up Robust Features," *J. Computer Vision and Image Understanding*, 110(3):346-359, 2008.
- [6] C.-H. Jan et al., "A 14 nm SoC Platform Technology Featuring 2<sup>nd</sup> Generation Tri-Gate Transistors, 70 nm Gate Pitch, 52 nm Metal Pitch, and 0.0499 μm<sup>2</sup> SRAM Cells, Optimized for Low Power, High Performance and High Density SoC Products," *IEEE Symp. on VLSI Tech.*, pp. 12-13, 2015.

**Figure 14.4.1: kNN accelerator organization.**

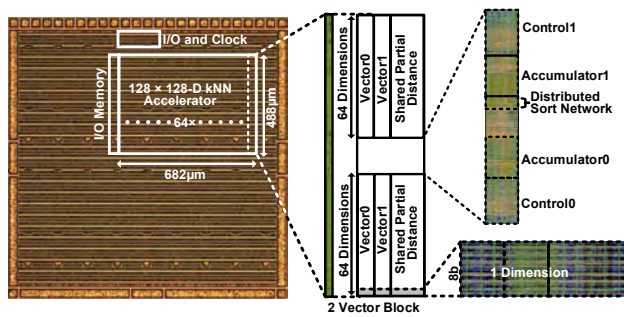
**Figure 14.4.2: Partial distance compute (a) iterations and (b) circuits.**

**Figure 14.4.3: Accumulator and sort circuits.**

**Figure 14.4.4: Search space control circuits.**

**Figure 14.4.5: 14nm CMOS measured throughput, latency, total power and energy vs. number of nearest neighbors computed (k).**

**Figure 14.4.6: 14nm CMOS measured throughput, total power and energy vs. supply voltage.**



|                                               |                                                                                              |
|-----------------------------------------------|----------------------------------------------------------------------------------------------|
| Process                                       | 14nm Tri-Gate CMOS                                                                           |
| Accelerator area, Transistor count            | 0.333mm <sup>2</sup> , 12.2M                                                                 |
| Nominal Performance, Power (k=1, 750mV, 25°C) | 21.5M vectors/s, 73mW (Manhattan), 1.8mW active leakage<br>20.3M vectors/s, 78mW (Euclidean) |
| Nominal energy efficiency (k=1, 750mV, 25°C)  | 3.37nJ/vector, 9.7TOPS/W (Manhattan)<br>3.84nJ/vector, 12.8TOPS/W (Euclidean)                |
| Peak performance                              | 26.4M vectors/s, 114mW, 850mV, 25°C (Manhattan, k=1)                                         |
| Peak efficiency (near-threshold)              | 1.23nJ/vector, 26.5TOPS/W, 2.1M vectors/s, 390mV, 25°C (Manhattan, k=1)                      |

**Figure 14.4.7: 14nm CMOS kNN accelerator die micrograph and measured performance summary.**