Skip to main content

> cs > arXiv:2010.11929

Search... GO

quick links

- Login
- Help Pages
- About

Computer Science > Computer Vision and Pattern Recognition

arXiv:2010.11929 (cs)

[Submitted on 22 Oct 2020 (v1), last revised 3 Jun 2021 (this version, v2)]

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

View PDF

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-

the-art convolutional networks while requiring substantially fewer computational resources to train.

Comments: Fine-tuning code and pre-trained models are available at this https URL. ICLR camera-ready version with 2 small modifications: 1) Added a discussion of CLS vs GAP classifier in the appendix, 2) Fixed an error in exaFLOPs computation in Figure 5 and Table 6 (relative performance of models is basically not affected)
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: **arXiv:2010.11929 [cs.CV]**
(or **arXiv:2010.11929v2 [cs.CV]** for this version)
https://doi.org/10.48550/arXiv.2010.11929
Focus to learn more
arXiv-issued DOI via DataCite

Submission history

From: Alexey Dosovitskiy [view email]
[v1] Thu, 22 Oct 2020 17:55:59 UTC (3,194 KB)
[v2] Thu, 3 Jun 2021 13:08:56 UTC (3,033 KB)

◉ Bibliographic Tools

Bibliographic and Citation Tools

☐ Bibliographic Explorer Toggle
Bibliographic Explorer *(What is the Explorer?)*
☐ Connected Papers Toggle
Connected Papers *(What is Connected Papers?)*
☐ Litmaps Toggle
Litmaps *(What is Litmaps?)*
☐ scite.ai Toggle
scite Smart Citations *(What are Smart Citations?)*

○ Code, Data, Media

Code, Data and Media Associated with this Article

☐ alphaXiv Toggle
alphaXiv *(What is alphaXiv?)*
☐ Links to Code Toggle
CatalyzeX Code Finder for Papers *(What is CatalyzeX?)*
☐ DagsHub Toggle
DagsHub *(What is DagsHub?)*
☐ GotitPub Toggle
Gotit.pub *(What is GotitPub?)*
☐ Huggingface Toggle
Hugging Face *(What is Huggingface?)*

☐ Links to Code Toggle
Papers with Code *(What is Papers with Code?)*
☐ ScienceCast Toggle
ScienceCast *(What is ScienceCast?)*
○ Demos

Demos

☐ Replicate Toggle
Replicate *(What is Replicate?)*
☐ Spaces Toggle
Hugging Face Spaces *(What is Spaces?)*
☐ Spaces Toggle
TXYZ.AI *(What is TXYZ.AI?)*
○ Related Papers

Recommenders and Search Tools

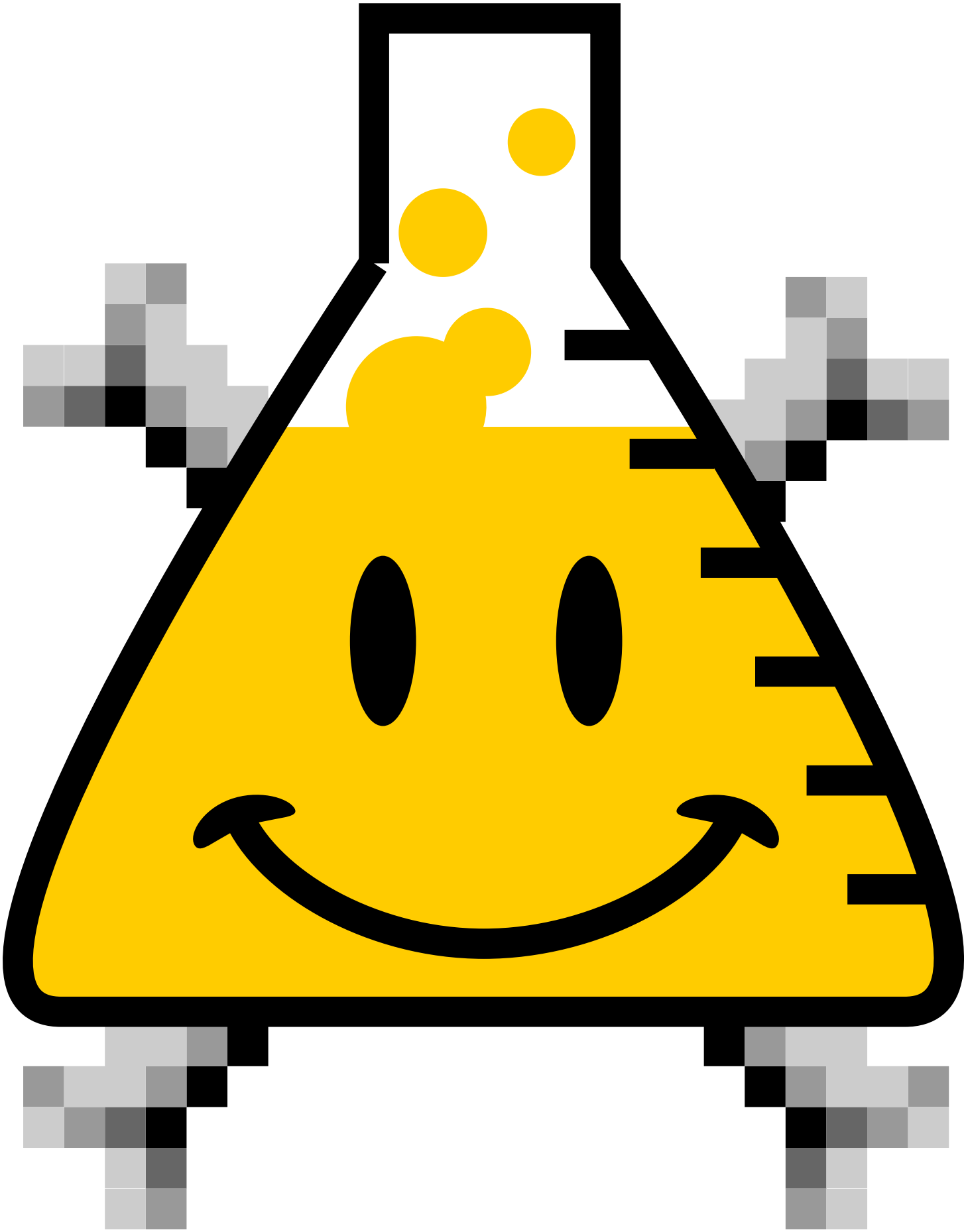
☐ Link to Influence Flower
Influence Flower *(What are Influence Flowers?)*
☐ Core recommender toggle
CORE Recommender *(What is CORE?)*
○ About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? **Learn more about arXivLabs**.



Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)