Transformers for Image Recognition at Scale

December 3, 2020

Posted by Neil Houlsby and Dirk Weissenborn,
Research Scientists, Google Research

Google Research

While convolutional neural networks (CNNs) have been used in computer vision since the 1980s, they were not at the forefront until 2012 when AlexNet surpassed the performance of contemporary state-of-the-art image recognition methods by a large margin. Two factors helped enable this breakthrough: (i) the availability of training sets like ImageNet, and (ii) the use of commoditized GPU hardware, which provided significantly more compute for training. As such, since 2012, CNNs have become the go-to model for vision tasks.

The benefit of using CNNs was that they avoided the need for hand-designed visual features, instead learning to perform tasks directly from data "end to end". However, while CNNs avoid hand-crafted feature-extraction, the architecture itself is designed specifically for images and can be computationally demanding. Looking forward to the next generation of scalable vision models, one might ask whether this domain-specific design is necessary, or if one could successfully leverage more domain agnostic and computationally efficient architectures to achieve state-of-the-art results.

As a first step in this direction, we present the Vision Transformer (ViT), a vision model based as closely as possible on the Transformer architecture originally designed for text-based tasks. ViT represents an input image as a sequence of image patches, similar to the sequence of word embeddings used when applying Transformers to text, and directly predicts class labels for the image. ViT demonstrates excellent performance when trained on sufficient data, outperforming a comparable state-of-the-art CNN with four times fewer computational resources. To foster additional research in this area, we have open-sourced both the code and models.

Google Research

The Vision Transformer treats an input image as a sequence of patches, akin to a series of word embeddings generated by a natural language processing (NLP) Transformer.

The Vision Transformer

The original text Transformer takes as input a sequence of words, which it then uses for classification, translation, or other NLP tasks. For ViT, we make the fewest possible modifications to the Transformer design to make it operate directly on images instead of words, and observe how much about image structure the model can learn on its own.

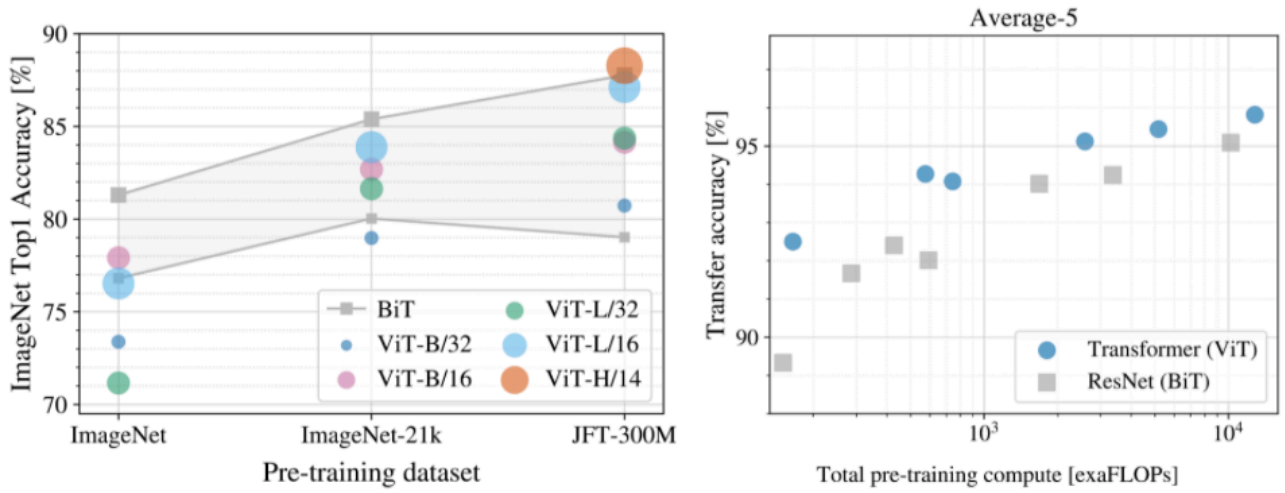
ViT divides an image into a grid of square patches. Each patch is flattened into a single vector by concatenating the channels of all pixels in a patch and then linearly projecting it to the desired input dimension. Because Transformers are agnostic to the structure of the input elements we add learnable position embeddings to each patch, which allow the model to learn about the structure of the images. *A priori*, ViT does not know about the relative location of patches in the image, or even that the image has a 2D structure — it must learn such relevant information from the training data and encode structural information in the position embeddings.

Scaling Up

mitigation strategies (e.g., regularization), ViT overfits the ImageNet task due to its lack of inbuilt knowledge about images.

To investigate the impact of dataset size on model performance, we train ViT on ImageNet-21k (14M images, 21k classes) and JFT (300M images, 18k classes), and compare the results to a state-of-the-art CNN, Big Transfer (BiT), trained on the same datasets. As previously observed, ViT performs significantly worse than the CNN equivalent (BiT) when trained on ImageNet (1M images). However, on ImageNet-21k (14M images) performance is comparable, and on JFT (300M images), ViT now outperforms BiT.

Finally, we investigate the impact of the amount of computation involved in training the models. For this, we train several different ViT models and CNNs on JFT. These models span a range of model sizes and training durations. As a result, they require varying amounts of compute for training. We observe that, for a given amount of compute, ViT yields better performance than the equivalent CNNs.



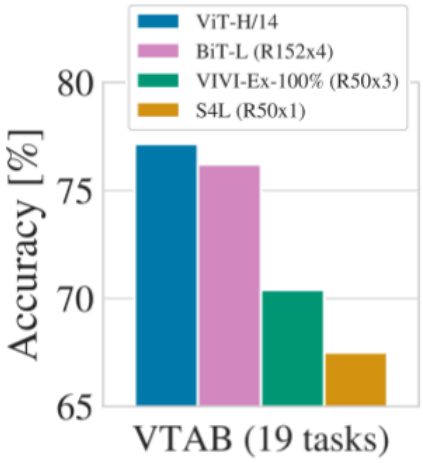
Left: *Performance of ViT when pre-trained on different datasets.* ***Right:*** *ViT yields a good performance/compute trade-off.*

High-Performing Large-Scale Image Recognition

Our data suggest that (1) with sufficient training ViT can perform very well, and (2) ViT yields an excellent performance/compute trade-off at both smaller and larger compute

This large ViT model attains state-of-the-art performance on multiple popular benchmarks, including 88.55% top-1 accuracy on ImageNet and 99.50% on CIFAR-10. ViT also performs well on the cleaned-up version of the ImageNet evaluations set "ImageNet-Real", attaining 90.72% top-1 accuracy. Finally, ViT works well on diverse tasks, even with few training data points. For example, on the VTAB-1k suite (19 tasks with 1,000 data points each), ViT attains 77.63%, significantly ahead of the single-model state of the art (SOTA) (76.3%), and even matching SOTA attained by an ensemble of multiple models (77.6%). Most importantly, these results are obtained using fewer compute resources compared to previous SOTA CNNs, e.g., 4x fewer than the pre-trained BiT models.

	ViT-H	Previous SOTA
ImageNet	88.55	88.5
ImageNet-ReaL	90.72	90.55
Cifar-10	99.50	99.37
Cifar-100	94.55	93.51
Pets	97.56	96.62
Flowers	99.68	99.63

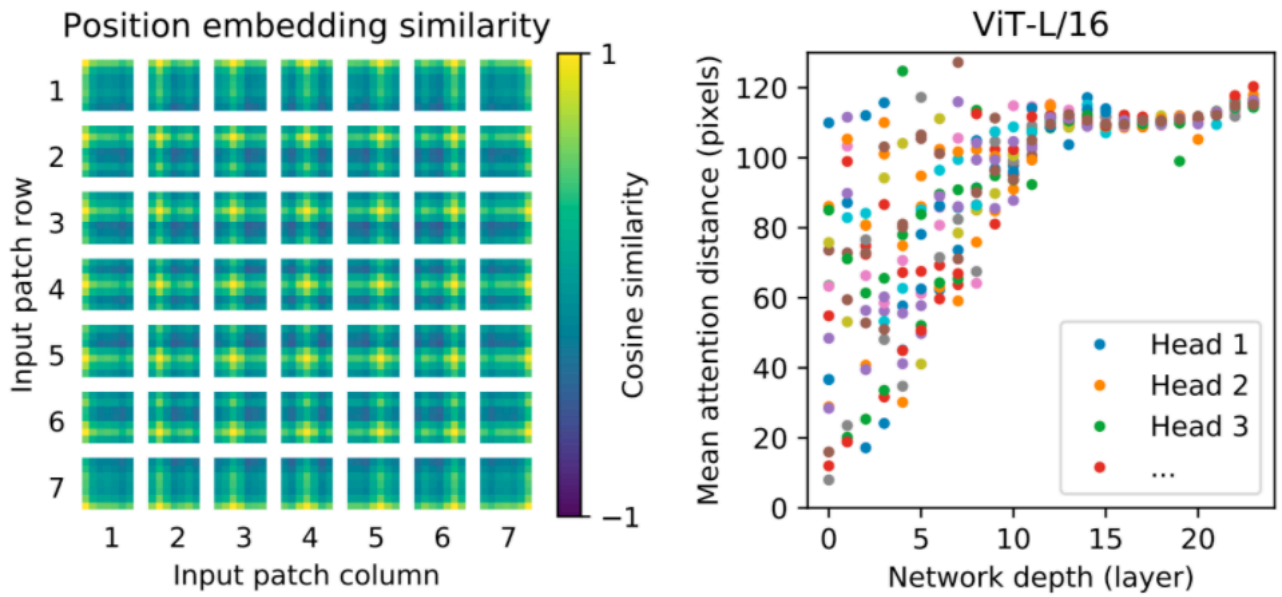


Vision Transformer matches or outperforms state-of-the-art CNNs on popular benchmarks. ***Left:*** *Popular image classification tasks (ImageNet, including new validation labels ReaL, and CIFAR, Pets, and Flowers).* ***Right:*** *Average across 19 tasks in the VTAB classification suite.*

Visualizations

To gain some intuition into what the model learns, we visualize some of its internal workings. First, we look at the position embeddings — parameters that the model learns to encode the relative location of patches — and find that ViT is able to reproduce an intuitive image structure. Each position embedding is most similar to others in the same row and column, indicating that the model has recovered the grid structure of the original images. Second, we examine the average spatial distance between one element attending to another for each transformer block. At higher layers (depths of 10-20) only global features are used (i.e., large attention distances), but the lower layers (depths 0-5) capture both global and local features, as indicated by a large range in the mean attention distance. By contrast, only local features are present

can aid generalization.



Left: *ViT learns the grid like structure of the image patches via its position embeddings.* ***Right:*** *The lower layers of ViT contain both global and local features, the higher layers contain only global features.*

Summary

While CNNs have revolutionized computer vision, our results indicate that models tailor-made for imaging tasks may be unnecessary, or even sub-optimal. With ever-increasing dataset sizes, and the continued development of unsupervised and semi-supervised methods, the development of new vision architectures that train more efficiently on these datasets becomes increasingly important. We believe ViT is a preliminary step towards generic, scalable architectures that can solve many vision tasks, or even tasks from many domains, and are excited for future developments.

A preprint of our work as well as code and models are publically available.

Acknowledgements

We would like to thank our co-authors in Berlin, Zürich, and Amsterdam: Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and Jakob Uszkoreit. We would like to thank Andreas Steiner for crucial help with infrastructure and open-sourcing, Joan Puigcerver and Maxim Neumann for work on large-scale

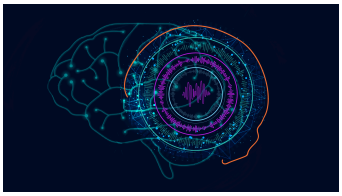
Google Research

Labels:

Machine Intelligence

Machine Perception

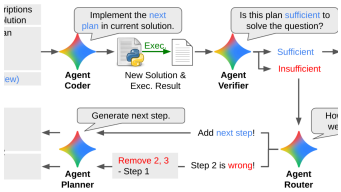
Other posts of interest



NOVEMBER 7, 2025

Introducing Nested Learning: A new ML paradigm for continual learning

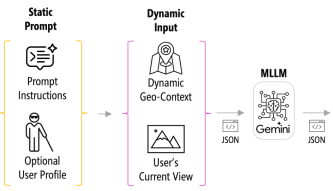
Algorithms & Theory · Generative AI · Machine Intelligence



NOVEMBER 6, 2025

DS-STAR: A state-of-the-art versatile data science agent

Data Mining & Modeling · Machine Intelligence · Natural Language Processing



OCTOBER 29, 2025

StreetReaderAI: Towards making street view accessible via context-aware multimodal AI

Generative AI · Human-Computer Interaction and Visualization · Machine Intelligence · Natural Language Processing