Review of *An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale*

Matthew Evans

May 7, 2025

Overview

In this paper[1], the authors explores how a Transformer model, traditionally used for text, can be repurposed for image recognition by dividing images into fixed-size patches treated as tokens. Unlike convolutional neural networks that embed spatial bias by design, the Vision Transformer must learn spatial layout and semantic relationships from data alone. Through large-scale pre-training and increased model capacity, the authors show that a pure Transformer can achieve performance on par with or exceeding state-of-the-art CNNs with minimal architectural changes.

Approach

Prior Work

Transformer Model [2] Introduced the Transformer model for machine translation based on multi-head self-attention and position-wise feed-forward layers, establishing the large-scale pre-training and fine-tuning paradigm later popularized by BERT[3] and GPT[4, 5].

Local Self-Attention [6, 7, 8, 9] Restricts self-attention to local neighborhoods and demonstrates that local multi-head attention blocks can completely replace convolutional layers in vision models.

Sparse Transformers [10] Employ scalable approximations to full self-attention—such as block-wise or strided patterns—to make global attention tractable on high-resolution images.

Block & Axial Attention [11, 12, 13] Explores applying self-attention along fixed blocks or individual tensor axes, reducing the quadratic cost of attention over flattened image patches.

Self-Attention & Convolution [14] Extract small 2×2 pixel patches and apply full self-attention, showing that a patch-based Transformer can match convolutional approaches on low-resolution images.

CNN-Attention Hybrids [15, 16, 17, 18, 19] Augment or post-process convolutional feature maps with self-attention for tasks such as classification, object detection, and video understanding.

iGPT [20] Apply a pure Transformer as an unsupervised generative model on reduced-resolution image pixels.

Novelty

The authors' key innovation is demonstrating that a *pure Transformer*, with only patch-based tokenization, a learnable classification token, and standard positional embeddings, suffices for state-of-the-art image recognition when pre-trained at scale, without the hand-crafted spatial biases of CNNs. By treating 16×16 or 32×32 image patches as tokens and reusing off-the-shelf Transformer encoders originally

designed for text, they show that "*large scale training trumps inductive bias*": large pre-training datasets (ImageNet-21k, JFT-300M) enable ViT to match or exceed CNN-based models with less compute.

Each input image $x \in \mathbb{R}^{H \times W \times C}$ (height, width, color channels) is reshaped into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$, flattened to vectors of dimension P^2C, and projected to a D-dimensional embedding. As in the NLP case, a learnable [`class`] token is prepended, and simple 1D positional embeddings are added (the authors note that 2D positional embeddings prove unnecessary). The resulting sequence is fed through L standard Transformer encoder layers (multi-head self-attention + MLP with GELU), and the final [`class`] output is linearly decoded for classification.

Patch-based Image Tokenization

By slicing images into fixed-size patches, ViT converts spatial data into a sequence suitable for Transformers. Larger patches (e.g., 16×16) reduce sequence length, balancing resolution and computation.

Minimal Vision-specific Inductive Bias

Unlike CNNs that intrinsically provide locality and translation equivariance at every layer, ViT's only bias is at the patch-extraction stage; all higher-order spatial relations are learned purely through attention.

Large-scale Pre-training

Pre-training on vast image corpora compensates for the lack of convolutional priors. ViT models pre-trained on JFT-300M achieve up to 88.55% ImageNet accuracy, outperforming comparably sized ResNets with substantially less pre-training compute.

Considerations

Strengths

- **State-of-the-art accuracy when pre-trained at scale:** ViT models pre-trained on large datasets (JFT-300M, ImageNet-21k) match or exceed top CNNs across ImageNet, CIFAR, and VTAB benchmarks.
- **Superior compute efficiency:** Vision Transformers use roughly 2-4× less pre-training FLOPs than comparably performing ResNets to reach the same downstream accuracies.
- **High memory-efficiency:** ViT can fit larger per-core batch sizes than ResNets at the same input resolution, aiding large-scale training.
- **Global receptive field from day one:** Self-attention layers integrate information across the entire image even in early layers, enabling direct modeling of long-range dependencies.

Weaknesses

- **Data-hungry and sample-inefficient on small datasets:** ViT overfits more than ResNets when trained on mid-sized subsets (e.g. 9M-30M), reinforcing that convolutional biases aid generalization under data scarcity.
- **Overfitting on limited data:** ViT overfits more than ResNets when trained on mid-sized subsets (e.g. 9M-30M), reinforcing that convolutional biases aid generalization under data scarcity.
- **Quadratic cost with image size:** Self-attention's bi-quadratic compute scaling in sequence length (i.e., number of patches) can become a bottleneck at very high resolutions.
- **High absolute pre-training cost for flagship models:** The largest ViT-H/14 variant requires thousands of TPU-core-days (e.g. 2.5k core-days for ViT-L/16) to pre-train, posing resource challenges.

Measures of Success

The authors assess quantitative success using standard classification metrics on downstream tasks. After fine-tuning, their largest ViT model pre-trained on JFT-300M achieves up to 88.55 % Top-1 accuracy on ImageNet, 90.72 % on ImageNet-Real, 94.55 % on CIFAR-100, and an average of 77.63 % across the 19-task VTAB suite. In few-shot linear evaluation—where representations remain frozen and only a linear classifier is trained, ViT surpasses comparably sized ResNets as pre-training data increases. Scaling studies further demonstrate that ViT reaches equivalent or better accuracies than ResNets with 2-4× fewer pre-training FLOPs.

The authors also present qualitative analyses of model behavior. By measuring *attention distance*, they show that some heads attend globally from the first layer while others focus locally, with average attention span expanding with depth, mirroring receptive field growth in CNNs. Attention rollout visualizations reveal that the `[class]` token predominantly attends to semantically meaningful regions (e.g., object parts), illustrating learned spatial and semantic grouping behaviors.

Impact

Evidence of the Vision Transformer's influence emerged quickly. Touvron et al. introduced the *Data-Efficient Image Transformer (DeiT)*[21], which incorporates a distillation token and teacher-student attention strategy to achieve 83.1% Top-1 accuracy on ImageNet-1K when trained from scratch on ImageNet in under three days on a single 8-GPU node. Liu et al. followed with the *Swin Transformer*[22], a hierarchical model using shifted non-overlapping windows to encode locality and multi-scale features, achieving 87.3% Top-1 accuracy on ImageNet-1K and demonstrating strong performance on detection and segmentation tasks. These works spurred a wave of subsequent architectures—including PVT[23], Twins[24], and CSWin[25]—that extend transformer efficiency, locality, and scalability across diverse vision tasks.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[6] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, and Alexander Ku. Image transformer. *CoRR*, abs/1802.05751, 2018.

[7] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. *CoRR*, abs/1904.11491, 2019.

[8] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *CoRR*, abs/1906.05909, 2019.

[9] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. *CoRR*, abs/2004.13621, 2020.

[10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019.

[11] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *CoRR*, abs/1906.02634, 2019.

[12] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *CoRR*, abs/1912.12180, 2019.

[13] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. *CoRR*, abs/2003.07853, 2020.

[14] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *CoRR*, abs/1911.03584, 2019.

[15] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *CoRR*, abs/1904.09925, 2019.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766, 2019.

[19] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020.

[20] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020.

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.

[23] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *CoRR*, abs/2102.12122, 2021.

[24] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. *CoRR*, abs/2104.13840, 2021.

[25] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *CoRR*, abs/2107.00652, 2021.