[Paper Review]
An Image is Worth 16x16 Words: Transformers For Image Recognition at Scale

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

”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.” [3]

1 Summary

ViTs were released as an alternative to convolutional neural networks (CNNs) by utilizing the fundamental concepts of transformers that have made them so successful in NLP settings to the task of vision. By reshaping the transformer model to handle 2D image patches, the authors of this paper develop a model which can leverage the strong efficiencies and capabilities of transformers. They find their architecture improves as model size and data size increases demonstrating that scale outperforms the inductive biases and equivariance of CNN architectures.

2 Introduction

Transformer based architectures have become the default choice for natural language processing. Vision modeling has been predominantly handled by CNNs. Inspired by the recent success of transformer architectures in NLP settings, the authors experiment with applications of modified transformers to the vision setting. They call this approach. ViTs work by splitting an image into multiple patches and treating each patch as a token exactly as seen in the NLP setting. They go on to demonstrate that ViTs are particularly efficient and good at significant scales. When trained on small datasets, ViTs under perform CNNs due to their lack of translation equivalence and locality; However, when scaled to larger datasets, ViTs significantly outperform models of comparable sizes across nearly every task. The simple take away is that transformer scaling capabilities outweigh the inductive bias of CNNs.

3 Background and Motivation

Transformers were designed as in context learning algorithms handling queries, keys, and values. It has since been adapted to the NLP setting. Direct application on a pixel-per-token equivalence would fail as input sizes scale and induce significant costs. Prior transformer vision approaches have approximated the locality of pixels similar to CNNs [7], but these have failed particularly when needing to handle global components. Instead, by utilize patches, as was also done in prior work to a lesser extent [2], the authors of this paper demonstrate the scalability of this patching approach to larger models and larger datasets. By applying this transformer approach to vision they demonstrate significant improvements over CNNs at scale.

3.1 Key Concepts

- **Concept 1:** In ViT, the equivalent tokens of NLP models are described as patches. Patches are two dimensional arrays which require simple modifications in the transformer architecture to enable training. Namely, reshaping the token process from $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = \frac{HW}{P^2}$ is the number of patches.
- **Concept 2:** ViTs work particularly well as data and model architecture is scaled to larger datasets. This is where the offer the most performance gains, both in memory efficiency, inference costs, and accuracy. Excluding the final MLP layers, ViTs lack of strong inductive bias or equivariance, but significantly outperform CNNs.
- **Concept 3:** The ViT architecture is simple, scalable, and efficient.

4 Methodology

To make this work, the authors reshape the token concept from the standard transformer to handle image 2D image patches rather than the traditional 1D token. Through reshaping the transformer outputs a projection as a patch embedding. This learnable embedding serves as the image representation that gets a classification head attached to each patch. This classification head is single hidden layer MLP for pre-training, and a single linear layer during fine-tuning. Finally, positional embeddings are introduced which provide a weak inductive bias and ultimately get sent as input to the transformer encoder. They also propose a hybrid model which replaces raw patches with feature maps from CNNs.

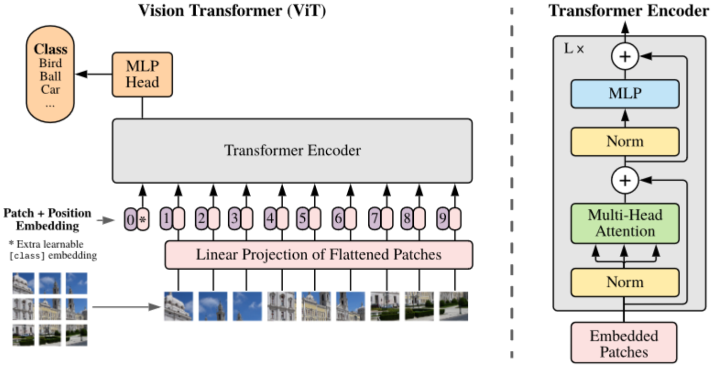


Figure 1: ViT Model Architecture. Splitting images into fixed patches, linearly embedding each patch, positional embedding, and adding a classification token to each patch to turn this into a classification task.

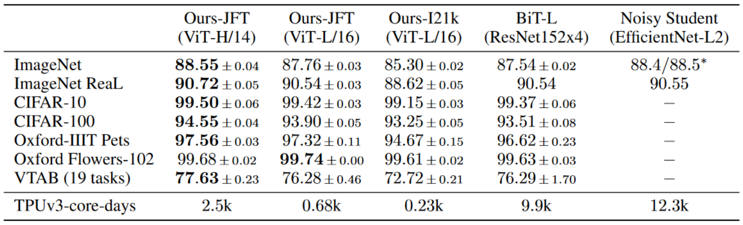
4.1 Overview of the Proposed Approach

- **Issue 1:** As described in depth, the transformer architecture handles 1D inputs and must be reshaped for 2D image patches. Extracting these patches are the only points where the model has access about the 2D image structure.

- **Issue 2:** This approach can handle arbitrarily long sequences with the consequence that position embeddings are no longer valuable. To counteract this, they interpolate the 2D interpolations according to their respective location in the given image.

5 Experiments and Results

In order to evaluate ViT, they compare the learning capabilities to ResNet and their hybrid model. They train on various datasets and evaluate their models on benchmark tasks like VTAB for classification. More specifically, they train on the ILSVRC-2012 ImageNet dataset which is comprised of 1k classes and 1.3M training images, ImageNet-21k with 21k classes, and 14M images, and JFT with 18k classes and 202M images. Their ResNet model is upgraded to utilize group normalization and standard convolutions to improve information transfer in order to provide a more fair comparison with ViT. This is because ViT skip connections are more powerful than those used in standard CNNs. They test their models on downstream datasets with few-shot or fine-tuned accuracies to capture the performance of each model. Through their experiments they demonstrate ViT significantly outperforms CNNs across every task and only improves as data is scaled. Notably, they train their models on TPUv3 for 30 days.



	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Figure 2: ViT comparison across a range of classification tasks.

5.1 Evaluation Metrics

- **Metric 1:** Models are evaluated utilizing a handful of benchmark tasks. ImageNet, ReaL labels [1], CIFAR-10/100 [4], Oxford-IIIT Pets [6], Oxford Flowers-102 [5], and the 19-task VTAB classification suite [8]. ”The tasks are divided into three groups: Natural – tasks like the above, Pets, CIFAR, etc. Specialized – medical and satellite imagery, and Structured – tasks that require geometric understanding like localization.” [3]
- **Metric 2:** Multiple ViT models are designed consisting of Base, Larger, and Huge. These not only define the model size, but the input patch size where Large defines a 16 × 16 image patch for the tokens. Each model is tested on its downstream performance on few-shot or fine-tuning accuracy’s formulated as a least-squares regression problem allowing for a closed form solution.

5.2 Key Results

- **Result 1:** Figure 2 illustrates the results of of their ViT models across a range of tasks. Outperforming ResNet and EfficientNet on every metric but a significant margin while taking significantly less computational resources to pre-train.
- **Result 2:** To explore pret-raining results, they test various data sizes with optimized hyper parameters and find that ViT outperforms its CNN counterparts only when the data size and model architecture is substantially large. When datasets and model architecture are small, ViTs overfit more than ResNets of comparable computational costs.
- **Result 3:** Models are pre-trained for $7 - 14$ epochs which result in ViTs outperforming ResNets on performance/compute trade-offs by using $2\times - 4\times$ less compute to achieve the same performance costs.
- **Result 4:** They find the internal representations of their positional embeddings encodes distance within the image patches. The attention mechanism also encodes global information within the early layers. Particularly, some attention heads become tuned to the task of global information while the other attention heads become tuned for localized attention. As model depth increases these attention distances increase.

6 Discussion and Critique

Despite the exceptional performance of ViTs over CNNs, their widespread adoption has been slow. This is likely due in part to the additional architectural complexity, and extensive datasets required for pre-training. CNNs are inherently efficient at capturing spatial hierarchies through their localized operations and transnational equivariance, properties which are baked into their design. ViTs, on the other hand, do not utilize inductive biases, relying instead on global self-attention mechanisms to overcome this difference. Furthermore, ViTs' reliance on a large-scale dataset during pre-training to outperform CNNs.

Additionally, the slow adoption of ViTs may also lie in the computational inefficiencies associated with their quadratic scaling in self-attention as the input size increases. Although ViTs eliminate the need for convolutions, their attention mechanism, which requires every patch to attend to every other patch, introduces scaling challenges, especially in high-resolution images.

6.1 Strengths

- **Strength 1:** ViTs scale significantly better as we increase model size and data set size. This results in better performance and accuracy over CNNs at the cost of minor increases in model complexity.
- **Strength 2:** ViTs are computationally efficient due to their transformer architecture improving both inferencing and energy costs over the long term.
- **Strength 3:** ViTs also handle small images patches well, by learning strong relative distance encodings between patches which leads to strong internal representations of global and localized patches.

6.2 Weaknesses

- **Weakness 1:** While ViTs show promise in benchmark settings, they require significant data resources for both training and fine-tuning.
- **Weakness 2:** ViTs lack the inductive biases present in CNNs, such as translation equivariance, which makes them less efficient at handling smaller datasets or tasks requiring strong spatial hierarchies. This limitation reduces their generalization capability in certain domains without extensive pre-training.

7 Future Directions

- Future work should explore introducing equivariance and stronger inductive bias into ViTs to improve their generalization capabilities on smaller datasets.
- Another direction of future work would be to explore strong model architecture decisions such as improved transformers or positional embeddings techniques.

8 Conclusion

Vits have emerged as a powerful alternative to CNNs at when done at scale. By leveraging the transformer, ViTs can capture long-range image dependencies for global and local information. Despite their impressive capabilities, strong performance at scale, and significantly reduced computational resources needed ViTs have not seen widespread adoption due to the necessary overhead to leverage their scaling capabilities. While they are likely the model of choice for scalable architectures, CNNs remain the strongest model architecture for their internal design choices leading to strong performance on small scale data and applications.

References

[1] Lucas Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *ArXiv*, abs/2006.07159, 2020.

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

[3] 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, 2021.

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[5] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[6] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[7] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018.

[8] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv: Computer Vision and Pattern Recognition*, 2019.