An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [1]

Lila Roig, Romane Barra

[1] 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.
[2] Implementing Vision Transformer (ViT) from Scratch, Tin Nguyen
[3] Training a Vision Transformer from scratch in less than 24 hours with 1 GPU, Saghar Irandoust, Thibaut Durand, Yunduz Rakhmangulova, Wenjie Zi, Hossein Hajimirsadeghi.

Introduction

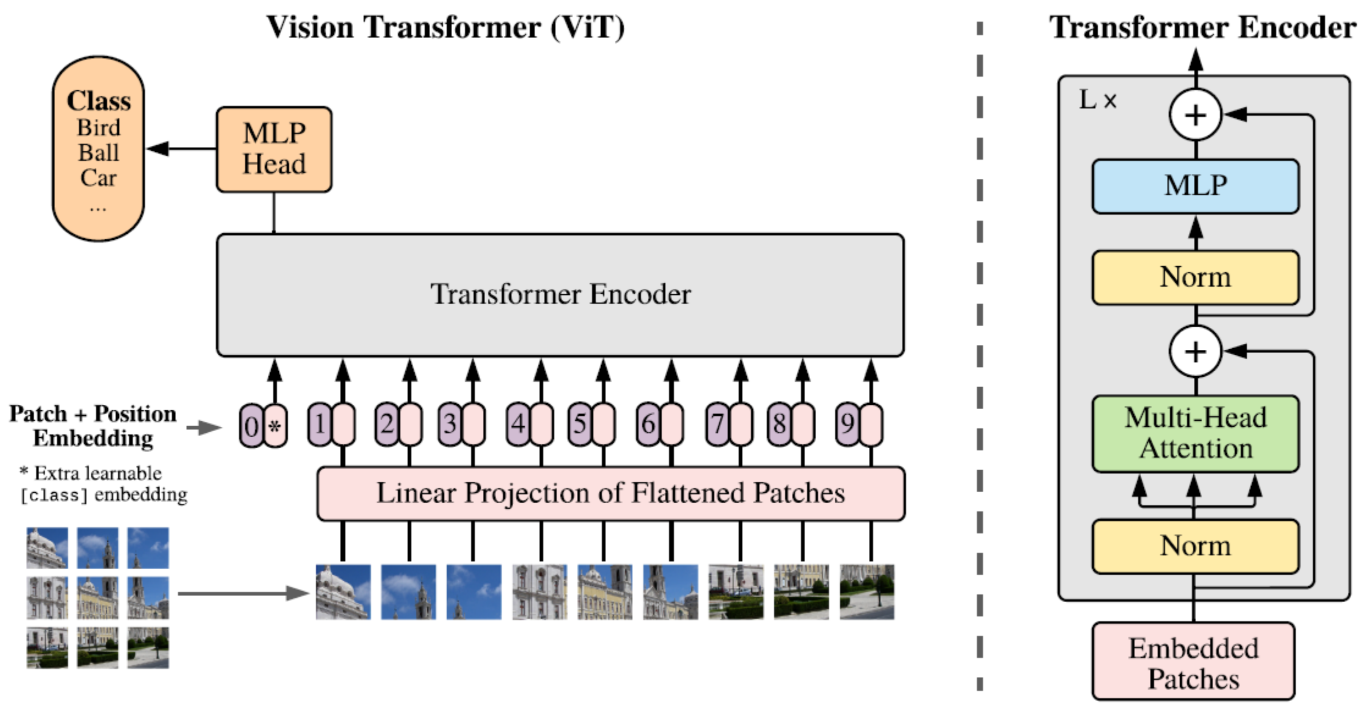
Deep learning has transformed computer vision, with **convolutional neural networks** (**CNNs**) excelling in tasks like image classification and object detection. Inspired by **Transformers**' success in NLP, this paper explores their use in vision tasks. The **Vision Transformer** (**ViT**) treats images as patch sequences, using self-attention to capture global dependencies, and shows promise in scaling and outperforming CNNs on large datasets.

Architecture

1. **Input image:** of size $H \times W \times C$.
2. **Divided into patches:** $N = \frac{HW}{P^2}$ patches of fixed size $P \times P$, then flattened into vectors of size $P^2 \times C$.

3. Linear projection: vectors are projected in a latent space of dimension D.

4. A Positional Embeddings is added to each vector to retain spatial information. **Special Classification Token** is prepended to the sequence. Its representation at the output is used for classification.



5. Transformer Encoder: The sequence of patch embeddings is processed by alternating layers of **Multi-Head Self-Attention** (**MSA**) and **Feed-Forward Networks** (**FFN**), with Layer Normalization and residual connections ensuring stability.

Self-attention calculates relationships between all patches by generating Queries Q, Keys K, and Values V from input embeddings.

The attention scores are computed as $S = \frac{QK^T}{\sqrt{D}}$, measures how much a source token "attends to" a target token.

The output for each query is : $Output = softmax(S)\, V$

6. Classification: The vector corresponding to the [*class*] token at the Transformer output is passed through a classification head : an MLP with one or more hidden layers, a simple linear layer during fine-tuning. The output is the scores for each class.

Datasets

To evaluate the scalability and performance of ViTs, the paper uses large-scale datasets:

- **ImageNet-21k**: 14 million images and 21,000 classes.
- **JFT-300M**: 300 million images and 18,000 classes.

For transfer learning and benchmarking, smaller datasets are also used: CIFAR-10/100, Oxford-IIIT Pets, Vision Task Adaptation Benchmark (VTAB).

In our code, we trained the model on the CIFAR-10 dataset.

Models

The paper introduces three variants of the Vision Transformer:

ViT-Base: 86M parameters, latent dimension 768, and 12 encoder layers.

ViT-Large: 307M parameters, latent dimension 1024, and 24 encoder layers.

ViT-Huge: 632M parameters, latent dimension 1280, and 32 encoder layers.

Results

Performance on Large-Scale Datasets:

- On ImageNet-21k, ViTs outperform ResNets and BiTs, demonstrating their ability to learn complex patterns.

- ViT-Huge achieves state-of-the-art performance on several tasks, such as 77.63% on VTAB (19 tasks) and 94.55% on CIFAR-100.

	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

Table 1 – Accuracy (with std) of ViTs compared to state of the art

Impact of Fine-Tuning: Fine-tuning on higher-resolution images allows ViTs to capture finer details, significantly boosting performance.

Importance of Pre-training: ViT learns robust feature representations from large datasets, improves its performance on smaller datasets and outperforms CNNs with the same computational budget.

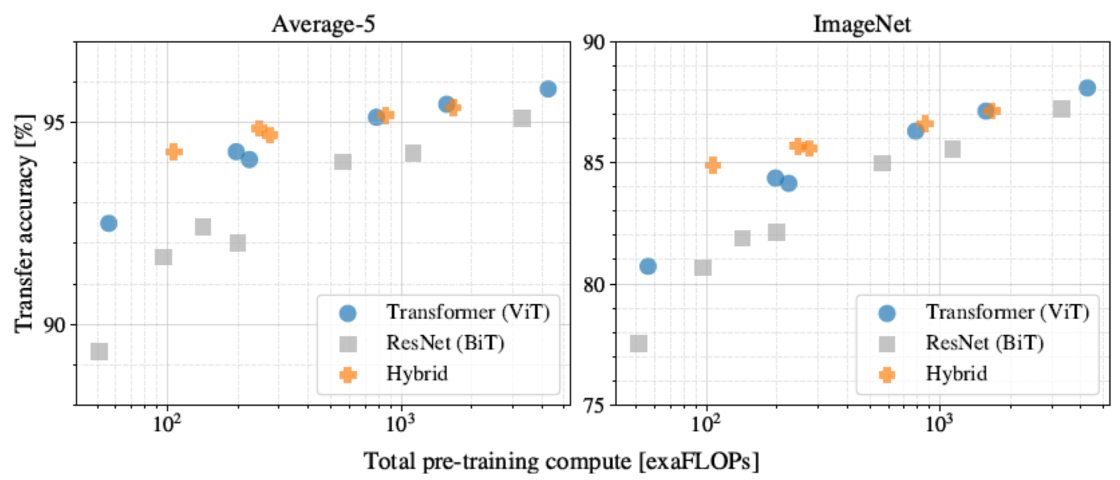


Figure 1 – Performance vs pre-training compute for different architectures

Personal experiments

We implemented our own ViT model and trained it from scratch on the small CIFAR-10 dataset [2]. After 40 epochs, completed in 27 minutes, the model achieved a train loss of 0.96, a test loss of 1.03, and an **accuracy of 0.63**. This performance is reasonable for a simplified ViT without pretraining. In contrast, the original ViT paper achieved near-perfect accuracy on CIFAR-10, largely due to extensive pretraining on large-scale datasets such as ImageNet-21k. This emphasizes the crucial role of pretraining in enabling ViTs to surpass CNNs on smaller datasets.

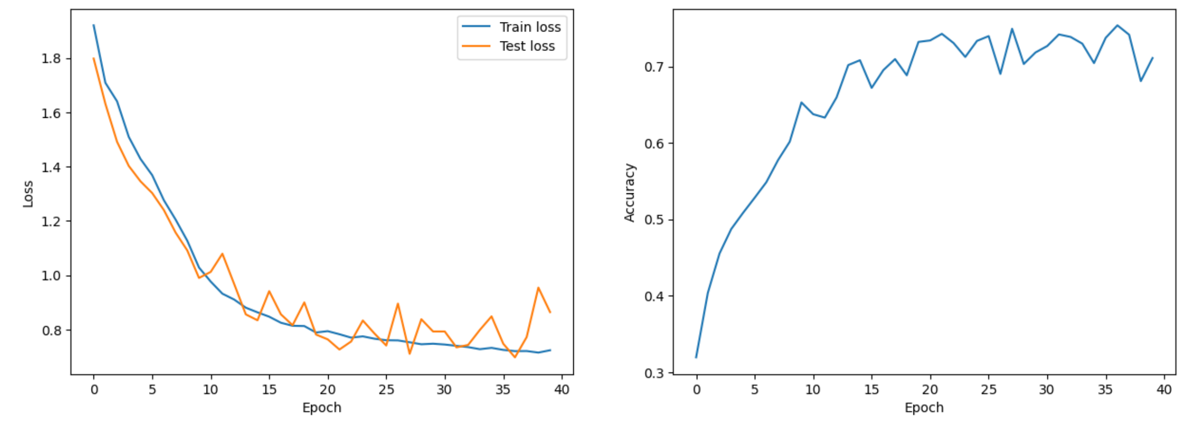


Figure 2 – Loss and accuracy during the training of our ViT.

We extended the code from [2] to include enhanced **visualization tools** to track how an image evolves through the model at key stages of training. We added an **MLP layer with locality** and implemented **curriculum learning**, gradually increasing the image size during training (recommendation in [3]).

These modifications aimed to accelerate convergence: after 40 epochs in 33 minutes, the model achieved a train loss of 0.72, test loss of 0.86, and **accuracy of 0.71**. While they reduced overfitting, computation was faster without them, suggesting greater benefits for larger datasets than small ones like CIFAR-10.

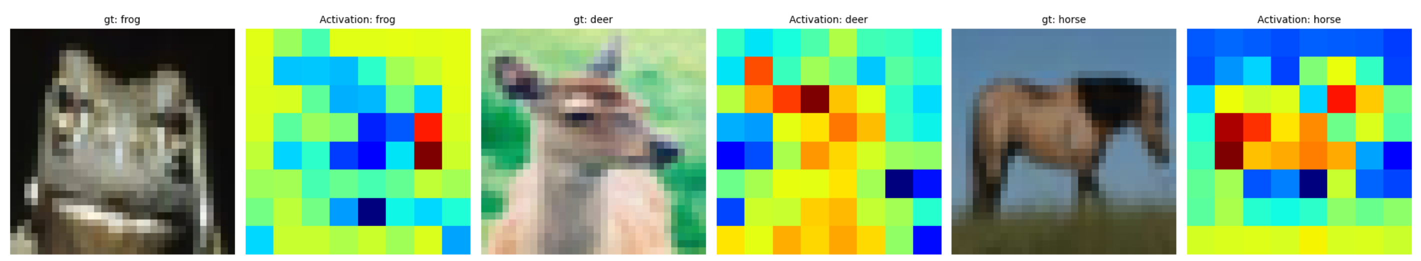


Figure 3 – Visualization of the patch embeddings

Figure 4 – Visualization of attention

Conclusion

The Vision Transformer (ViT) redefines computer vision by leveraging self-attention to **achieve state-of-the-art performance**. Future directions include **hybrid models** combining CNNs and Transformers and developing efficient self-attention mechanisms to **reduce computational costs** and broaden applicability, like high-resolution images and real-time tasks.