An Image is Still Worth 16x16 Words

Akaash Mahinth, Cody Torgovnik, Daniel Lines
github.com/ctorgovnik/ViT-Project

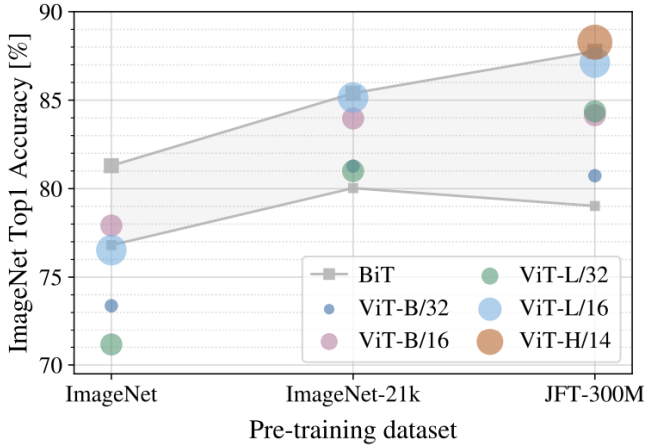
1 Introduction

Following the 2017 paper *"Attention is All You Need"* by Vaswani et al.[1], the Transformer architecture quickly became foundational in the machine learning landscape. Researchers in computer vision began to ask: Can the Transformer be applied effectively to images for large-scale recognition tasks? This question was addressed by Dosovitskiy et al. from Google Research's Brain Team in their 2020 paper, *"An Image is Worth 16×16 Words."*[2] In this seminal work, they introduced a novel method for tokenizing images into fixed-size patches, enabling the use of the Transformer's attention mechanism on visual data. The resulting Vision Transformer (ViT) model achieved state-of-the-art results on several image classification benchmarks, demonstrating that attention based models could surpass convolutional architectures at scale. Key contributions of the paper include the introduction of the ViT architecture, a new image embedding strategy using patch tokenization, and empirical evidence of ViT's superior performance when trained on large datasets.

2 Chosen Results

We chose to reproduce the scaling results presented by Dosovitskiy et al. in their Vision Transformer (ViT) paper[2]. Specifically, our goal was to replicate the findings that demonstrate how convolutional neural networks (CNNs), particularly ResNets, outperform ViTs when trained on smaller datasets due to the strong inductive biases inherent in convolutional architectures. Conversely, we aimed to show that as dataset size increases, these inductive biases become less critical, and ViTs begin to outperform CNNs and ResNets, leveraging their higher model capacity and scalability. Looking at figure 1, pulled directly from *"An Image is Worth 16×16 Words."*, we first notice that when comparing the architectures trained on ImageNet, the smallest of the datasets presented, the top of the CNN accuracy range (gray band) is well above any of the ViT models that they tested in the paper. But when moving to much larger datasets, such as the full ImageNet or especially the massive JFT-300M dataset, the ViTs jump in performance. The finetune accuracy of the larger ViTs trained on JFT-300M, for example, are at the top of the band for CNN accuracy, and the largest ViT model they tested out performed the best CNN style models of the time. This data shows how when access to large datasets is not an issue, ViTs can significantly outperform convolutional methods, which redefined what top of the line image classification models look like.

Figure 1: Graph comparing different pretrain corpuses with fine tune accuracy. Larger datasets result in better finetune for ViTs.[2]



3 Methodology

As pointed out in the paper, training ViTs is very resource intensive. They are much more compute heavy then convolutional models, which is one of their drawbacks. For us, as undergraduate students without access to large compute, this meant we would have to find another way to test the scaling of these models. We decided to mix models designed by ourselves with larger models pulled from resources on the internet.

3.1 Models

We utilized a few models to illustrate the scaling law that we wanted to reproduce. Our chosen CNN was the ResNet, as described below. We used 3 different ViTs for our tests, from an in house model (very small, easy to train), to much larger pretrained models. Pretraining is described below, and all models were finetuned and tested on CIFAR-10[3].

- **In House ViT**: We implemented a ViT as described in the paper[2], with approximately 1M parameters. We pretrained on CIFAR-100[3].
- **Mini ViT**[4]: We used Facebook's deit-tiny-patch16-224 to simulate a very small ViT pretrained on ImageNet1K. It has approximately 5M parameters.
- **Base ViT**[5]: The base ViT model as described in the paper. This model was pulled from PyTorch, and is pretrained on ImageNet1K.
- **ResNet-18**[5]: A prebuilt ResNet from PyTorch, approximately 11.6M parameters. We pretrained a version of this model on CIFAR-100[3], and pulled another version pretrained on ImageNet1K.

3.2 Datasets

We utilized 3 main data sets for our findings. Our finetuning dataset was CIFAR-10[3], consistently used across all models and pretraining sizes. For our small pretraining dataset, we used CIFAR-100[3], a corpus of about 60000 images divided into 100 classes. Finally for our big pretraining dataset, we used ImageNet1k. We didn't run the pretraining ourselves, as this is a massive dataset and we did not have the compute to do this, but all models pulled from PyTorch and Huggingface were pretrained on this dataset.

3.3 Evaulation Metrics and Modifications

As mentioned above, we evaluated each model by finetuning on CIFAR-10[3] and analyzing the finetuned accuracy. We are looking for differences in finetune accuracy depending on the pretrain dataset size. Because of our limited access to compute, the main difference from our methodology to the papers is that everything is scaled down, from the size of the models to the size of the pretraining data sets. Our biggest model is about seven times smaller than the largest model used in the paper, and our largest pretraining dataset (ImageNet1K) is about 300 times smaller than the largest dataset they used (JFT-300M).

4 Results and Analysis

Table 1 displays the statistics we found during our research. The top section of the table shows the statistics for our small in-house ViT vs the ResNet-18, both pretrained on CIFAR100[3] and then finetuned and tested on CIFAR10[3]. As can be seen in the table, the convolutional model was able to outperform the ViT in this low data environment. In the bottom half of table 1, you can see the performance of the Mini ViT, Base ViT, and once again ResNet-18[5] but this time pretrained on the ImageNet-1K dataset. With this much larger pretraining dataset, the ViTs were able to achieve much higher finetune accuracy, and in the case of the Mini ViT vs ResNet, higher accuracy even with fewer parameters.

Model	# Parameters	Pretraining Dataset	CIFAR-10 Accuracy
ViT (OC)	1M	CIFAR-100	65.00%
ResNet-18	11.6M	CIFAR-100	72.36%
Mini ViT	5M	ImageNet-1K	**87.33%**
Base ViT	86.5M	ImageNet-1K	**94.64%**
ResNet-18	11.6M	ImageNet-1K	79.73%

Table 1: Comparison of model parameters, pretraining datasets, and CIFAR-10 accuracy.

In figure 2, the trend of ViTs performing worse than ResNets under small datasets but outperforming ResNets with large sets is clearly shown. Additionally, the rate of change between the small and large dataset of each model indicates future performance of the model as the pretraining dataset size increases. The ResNet's accuracy increased from 72% to 79% whereas the ViT's accuracy increased from 65% to 87.33% (under the mini ViT). This difference implies that the ResNet's performance is plateauing and will continue to be out performed by the ViT as the pretraining dataset increases in size.

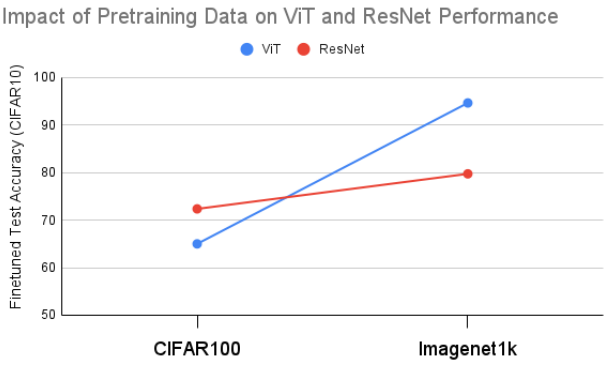


Figure 2: Graph of VIT and ResNet performance on CIFAR-10 depending on pretraining dataset

Our findings support the results found in the original paper, showing that CNNs inductive bias is very powerful for small datasets, letting the convolutional models learn much quicker and with less compute. But when access to large compute and large datasets is not an issue, the ViTs set new benchmarks for image classification tasks. In the larger CV space, these findings brought light to a new architecture that would go on to be considered state of the art for image classification, especially with the readily available datasets like ImageNet and JFT. And finally in the larger ML space, a viable way to implement the Transformer architecture for images had been found, something that many researchers had been interested in doing since the inception of the attention mechanism. The use of the class embedding, the use of a positional embedding, and the flattening of patches were combined a novel technique that would change the CV landscape for years to come.

5 Conclusion and Future Work

Our experiments reproduced the core finding of Dosovitskiy et al., showing that ViTs require large-scale pretraining to outperform convolutional models like ResNet. On smaller datasets such as CIFAR-100, ResNet-18 showed stronger performance, but ViTs significantly outperformed when pretrained on ImageNet-1K. These results highlight the scalability of ViTs and their potential when ample data and compute are available.

In future work, we aim to investigate methods that reduce ViTs' data and compute requirements, such as knowledge distillation, data augmentation, or semi-supervised learning. We could also explore hybrid architectures that combine convolutional and attention-based components to improve performance on smaller datasets.

References

[1] Ashish Vaswani et al. “Attention is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[2] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations (ICLR)*. arXiv:2010.11929. 2021.

[3] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[4] Facebook AI. *deit-tiny-patch16-224 on HuggingFace*. https://huggingface.co/facebook/deit-tiny-patch16-224. 2024.

[5] PyTorch Team. *Torchvision Models*. https://pytorch.org/vision/main/models.html. 2024.