PRE-REPORT: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

SongSeungWu

1 ABSTRACT

The Transformer architecture has become the standard for natural language processing tasks but remains underutilized in computer vision. This paper demonstrates that a pure Transformer, applied directly to sequences of image patches, performs effectively in image classification tasks. Vision Transformer, pre-trained on large datasets, achieves excellent results on benchmarks like ImageNet, CIFAR-100, and VTAB, outperforming state-of-the-art CNNs while requiring fewer computational resources.

2 INTRODUCTION

Self-attention-based architectures, particularly Transformers, have become standard in NLP due to their scalability. In contrast, CNNs remain dominant in computer vision. This paper applies standard Transformers to images by splitting them into patches, treating patches as tokens in NLP tasks. While initial results on smaller datasets showed limitations, large-scale pretraining on datasets like JFT-300M revealed ViT's ability to surpass CNNs in benchmarks such as ImageNet (88.55%) and CIFAR-100 (94.55%).

3 RELATED WORK

Transformers have been successfully applied in NLP, and various methods have been explored to adapt them for image processing. Approaches like local Self-attention or Sparse Transformers reduce computational cost but require complex hardware implementation. Cordonnier et al. proposed using 2×2 patches for Self-attention, but ViT handles larger patches and medium-resolution images, demonstrating superior performance over CNNs when pre-trained on large datasets. Generative approaches like iGPT also exist but underperform compared to ViT in image classification. This paper explores the transfer learning capabilities of Transformers on large-scale datasets like ImageNet-21k and JFT-300M, comparing them to CNN-based models.

4 METHOD

- **Vision Transformer**: ViT processes images by reshaping them into patches and converting them into a 1D sequence with learnable position embeddings and a class token. The Transformer encoder uses self-attention and MLP blocks. Unlike CNNs, ViT relies on training to learn spatial relationships rather than inherent inductive biases.

- **Fine-tuning and Higher Resolution**: Pre-trained ViT models are fine-tuned for smaller tasks. Higher-resolution fine-tuning benefits performance by increasing sequence length, with position embeddings adjusted using 2D interpolation.

5 EXPERIMENTS

The representation learning capabilities of Vision Transformer, ResNet, and hybrid models were evaluated using datasets of varying sizes and benchmarks. ViT achieved state-of-the-art performance on most recognition benchmarks while maintaining lower pre-training costs. Additionally, self-supervised ViT demonstrated promising potential for future applications.

- **SETUP**:

  - **Datasets and Setup**: The evaluation included large-scale datasets like ImageNet, ImageNet-21k, and JFT, as well as benchmarks like CIFAR-10/100 and VTAB. VTAB assesses low-data transfer capabilities and categorizes tasks into Natural, Specialized, and Structured groups.

  - **Model Variants**: ViT configurations are based on BERT, with Base, Large, and Huge variants depending on model size and patch size. ResNet-based CNNs used modified "ResNet" structures incorporating Group Normalization and standardized convolutions.

  - **Training and Fine-tuning**: All models were trained with the Adam optimizer and high weight decay using linear learning rate warmup and decay. Fine-tuning employed SGD with momentum, and ViT achieved optimal performance by fine-tuning at higher resolutions.

  - **Metrics**: Models were evaluated using Fine-tuning and Few-shot accuracies. Fine-tuning measures model performance after task-specific adjustment, while Few-shot evaluates representations via regularized least-squares regression, offering a computationally efficient alternative for quick assessments.

- **Comparison to State-of-the-Art**: Large Vision Transformer models outperform ResNet-based Big Transfer and EfficientNet-based Noisy Student in most benchmarks while requiring lower pre-training costs. Particularly, ViT-H/14 pre-trained on JFT-300M shows superior performance on datasets like ImageNet, CIFAR-100, and VTAB. ViT tends to surpass CNNs as the size of the pre-training dataset increases.

- **Pre-training Data Requirements**: ViT excels on large datasets like JFT-300M but underperforms on smaller datasets where CNNs have the advantage due to inductive biases. While ViT is prone to overfitting on small datasets, it benefits from directly learning patterns from large datasets. For instance, ResNet outperforms ViT on a 9M dataset, but ViT surpasses on datasets with 90M+ images.

- **Inspecting Vision Transformer**: ViT projects flattened patches into a lower-dimensional space and adds position embeddings to learn spatial relationships between patches. The position embeddings represent 2D topology, with closer patches having more similar embeddings. Self-attention integrates global information even in early layers, while some attention heads focus on localized details, functioning similarly to early CNN convolutional layers. As the depth increases, the attention range expands, focusing on semantically relevant image regions for classification.

6 CONCLUSION

ViT applies Transformers to image recognition with minimal modifications, achieving competitive or superior performance to CNNs on various datasets. While promising, challenges remain in extending ViT to tasks like object detection and segmentation and further optimizing self-supervised learning. Scaling ViT further is expected to yield even better performance.