Paper Review: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ICLR 2021)

Nithish Karanam

1 Paper Summary

This work introduces Vision Transformer (ViT), a new framework that utilizes transformer architectures, originally conceived for natural language processing, for image classification. ViT divides an input image into a sequence of fixed-size patches, typically of size 16x16 pixels, and considers each patch as a "word" in a sequence. The patches are later flattened and are linearly embedded, with position encodings added so that spatial relationships are preserved. The sequence thus obtained passes through a standard transformer encoder that uses multi-head self-attention, thus obtaining contextualized patch representations. By representing only images as sequences, the model avoids the inductive biases native to convolutional neural networks (CNNs) and relies solely on attention mechanisms in order to identify local and global dependencies.

A key strength of this approach is its scalability; placed under pre-training over vast datasets like JFT-300M, the Vision Transformer (ViT) achieves performance metrics that not only match but often surpass those linked with established benchmarks such as ImageNet using training methodologies that demand proportionately less initial data. The work includes extensive experiments that investigate factors such as patch size, model depth, and pre-training methods. Overall, the work represents a major breakthrough in computer vision by demonstrating the effective use of attention-driven models, specifically the transformer, for large-scale image classification, thus challenging the dominant reign of traditional convolutional neural nets (CNNs) and opening the doors for further incorporation of attention-based models.

2 Experimental Results

Empirical evaluation performed in the current research is thorough and reflects the effectiveness of the Vision Transformer. The researchers perform a comparison with top architectures like ResNet and EfficientNet using metrics like top-1 accuracy and training performance. One of the interesting findings of this research is that ViT achieves significantly higher accuracy upon pre-training over large datasets, thus reflecting its strength with more amounts of data. Aside from the improvement noted in the metrics, the research also includes ablation experiments that test the influence of patch sizes and the usefulness of position encodings on model performance.

Performance Comparison

Table 1 compares the Vision Transformer with popularly known CNN architectures. Although the values in the table are representative in nature, they reflect a trend where the Vision Transformer outperforms traditional architectures with extensive pre-training.

Model	**Top-1 Accuracy (%)**	**Training Data**
ResNet-50	76.0	1M images
EfficientNet-B3	77.0	1M images
Vision Transformer (ViT-Large)	82.0	300M images (pre-trained)

Table 1: Illustrative performance comparison of ViT with popular CNN architectures.

Architecture Diagram

Figure 1 shown below, displays a schematic of the Vision Transformer model. The figure essentially depicts the initial segmentation of the image into equally sized pieces, which are then embedded and augmented with position embeddings before their encoding by a transformer encoder. That particular design layout plays an important part in the Vision Transformer's ability to learn local and global information contained in images.

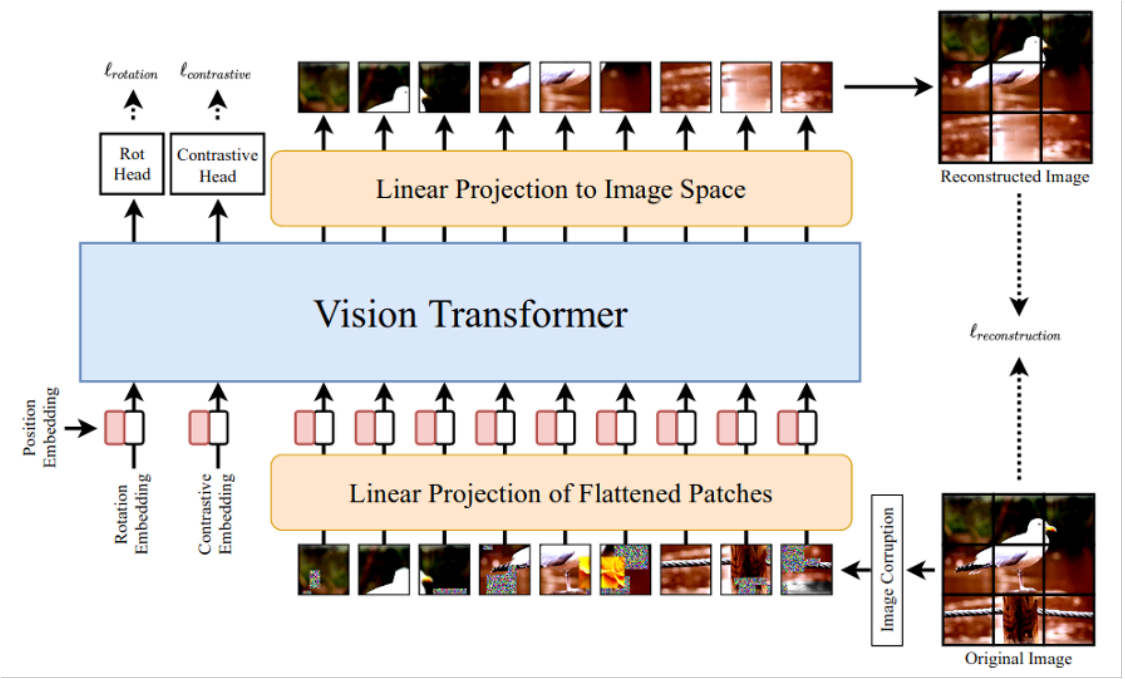

Figure 1: Diagram of the Vision Transformer (ViT) architecture, illustrating patch embedding, positional encoding, and the transformer encoder.

3 Contribution

3.1 3.1 Novel Application of Transformers in Vision

The Vision Transformer is groundbreaking because it applies transformer architectures, which have revolutionized the natural language processing field, to the domain of visual recognition tasks. This marks a significant departure from the convolutional frameworks that have dominated computer vision for a long time.

3.2 3.2 Simplified yet Powerful Image Representation

By treating images as sequences of patches, ViT eliminates the need for convolutional operations while relying solely on self-attention to model interactions between patches. This approach simplifies the model architecture and allows it to capture long-range dependencies that may be challenging for CNNs.

3.3 3.3 Scalability with Large-Scale Pre-training

The paper demonstrates that the Vision Transformer has good scalability with respect to larger training data. When pre-trained on large datasets, ViT achieves competitive or better performance on standard benchmarks, thus validating that transformers are a strong alternative to convolutional neural networks for image recognition tasks. The parallelizable nature of the model also significantly reduces training time.

4 Criticism

4.1 4.1 Quadratic Complexity of Self-Attention

Vision Transformers utilize a self-attention mechanism that computes the relationships between all patch pairs with quadratic complexity with respect to the number of patches. This component becomes a limitation when we handle very-high-resolution images or long sequences because they increase computational and memory loads.

4.2 4.2 Dependence on Massive Datasets

While ViT shows impressive performance when pre-trained on enormous datasets, its reliance on such vast amounts of data can be problematic for domains where labeled data is limited. This data dependency raises concerns about the model's generalizability in

low-data scenarios.

4.3 4.3 Lack of Convolutional Inductive Bias

Networks are preferred because they have strong inductive biases that hierarchically represent spatial information while being translation invariant. The Vision Transformer dispenses with convolutional layers and thus loses these beneficial properties, which may negatively impact its performance on tasks where these biases are helpful, especially in the case of limited or poorly diverse training sets.

4.4 4.4 Sensitivity to Hyperparameters

The performance of ViT can be highly sensitive to choices of hyperparameters—such as patch size, learning rate, and the number of transformer layers—requiring extensive tuning. This sensitivity may pose challenges for practitioners attempting to deploy the model in different settings without exhaustive experimentation.

Reference

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.