# Transformer for image recognition pre-report

Taehun Kim[1]
[1] Department of Computer Science and Engineering, Pusan National University.

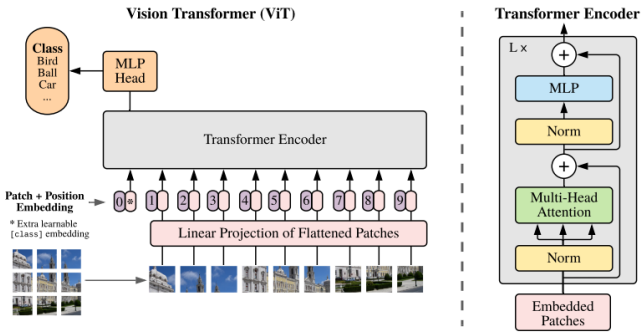


Figure 1: Vision Transformer model overview.

## 1 Introduction

Transformer[3] architecture is the model of choice in natural language processing. In computer vision, however, Convolutional Neural Network is dominant. Due to the success of Transformer, there is some works try to replace some convolutional layers to self-attention.

On the other hand, Vision Transformer[1] apply standard Transformer directly to images, without any Convolutional layers. This model achieves excellent results in transfer learning based on pretraining with a large dataset(e.g. ImageNet-21k).

In this report, We summarize Vision Transformer paper[1] focused on its architecture.

## 2 Multihead self-attention

self-attention generate Key, Value, Query vector from input vector. the size of matrix $U$ is $(D \times 3D_k)$ when an input sequence is $(N \times D)$

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv}$$

The attention weights is similarity between query and key.

$$A = \texttt{softmax}(\mathbf{q}\mathbf{k}^T / \sqrt{D_h})$$

Output vector is product of attention weights and value.

$$SA(\mathbf{z}) = A\mathbf{v}$$

Multihead self-attention(MSA) is running $k$ self-attention operations, combining them, and apply linear projection.

$$MSA(\mathbf{z}) = [SA_1(z); SA(z); \cdots; SA_k(z)]\mathbf{U}_{msa}$$

## 3 Vision Transformer

model architecture is shown in Figure . The image was reshaped from $(H \times W \times C)$ into $(N \times (P^2 \cdot C))$, where H, W, and C is height, width, and channel of the image, $(P \times P)$ is resolution of patch, and N is the number of patches. A flattened patch passes through a D-dimensional vector to become a patch embedding vectors.

Position embeddings are added to the patch embedding. They use standard learnable 1D vector for position embeddings.

Additionally, There are extra learnable class embedding similar to BERT's class token.

| Model | Layers | Hidden D | MLP size | Heads | Params |
|---|---|---|---|---|---|
| ViT-Base | 12 | 768 | 3072 | 12 | 86M |
| ViT-Large | 24 | 1024 | 4096 | 16 | 307M |
| ViT-Huge | 32 | 1280 | 5120 | 16 | 632M |

Table 1: Details of Vision Transformer model variants.

The Transformer encoder consists of multiheaded self-attention(MSA) and MLP blocks. Layernorm(LN) is applied before every block.

$$\mathbf{z}_0 = [\mathbf{x}_{class}; \mathbf{x}_p^1 E; \mathbf{x}_p^2 E; \cdots \mathbf{x}_p^N E;] + E_{pos}$$

$$\mathbf{z}'_l = MSA(LN(\mathbf{z}_{l-1})) + z_{l-1}$$

$$\mathbf{z}_l = MLP(LN(\mathbf{z}_{l-1})) + z_l$$

$$y = LN(\mathbf{z}_L^0)$$

Transformer has much lass inductive bias than CNNs. only MLP layers are local and translationally equivariant.

They pretrain Vision Transformer using large datasets, and fine tune to smaller tasks. this network can handle arbitrary sequence length. however, when fine-tuning with higher resolution image than pre-training, position embeddings may no longer meaningful. They perform interpolation to the position embedding.

The details of the model variants are shown in Table 1. For instance, ViT-L/16 means the large variant with $16 \times 16$ patch size. Large and Huge model pretrained with large dataset(JFT, Imagenet 21k) achieve better performance in popular image classification benchmarks(e.g. ImageNet, CIFAR-100, VTAB) while took less compute to pre-train than state-of-the art.

### 3.1 Pre-training data requirements

Transformer[3] is modest when pretrained with smaller dataset due to have fewer inductive biases than CNN. however, ViT performs better when pretrained with large dataset such as JFM-300M than ResNet. Additionally, ViT uses approximately $2-4\times$ less compute to attain the same performance.

## 4 Conclusion

ViT is direct application of Transformer architectures to image classification. This model achieve better performance when pretrained with large dataset than CNN model or Big Transfer[2], Additionally, this model is cheaper to pre-train.

[1] 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. URL https://arxiv.org/abs/2010.11929.

[2] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2020. URL https://arxiv.org/abs/1912.11370.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.