Lab 5: Vision transformer pre-report

CHI YEONG HEO[1],
[1]School of Computer Science and Engineering, Pusan National University, Busan 46241 Republic of Korea

1. Introduction

This report provides a summary of a paper in the field of computer vision: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [1].

2. Overview of Vision Transformer

The Vision Transformer (ViT) is a transformer architecture applied to image classification, originally designed for natural language processing. Unlike convolutional neural networks (CNNs), ViT processes images by dividing them into fixed-size patches. Each patch is flattened into a 1D vector, allowing the model to treat the image as a sequence of tokens, similar to words in a sentence.

3. Vision Transformer (ViT)

The standard Transformer architecture is designed to process 1D sequences of token embeddings. In adapting this structure for 2D images, an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is partitioned into a sequence of flattened 2D patches, denoted as $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Here, (H,W) represents the resolution of the original image, C is the number of channels, (P,P) is the resolution of each individual image patch, and $N = HW/P^2$ defines the total number of patches, effectively setting the length of the input sequence for the Transformer. Each patch is flattened and then mapped into a D-dimensional space through a trainable linear projection (Eq. 1).

Similar to the `[class]` token in BERT, a learnable embedding $\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$ is prepended to the sequence of embedded patches, serving as the image representation $\mathbf{y}$ obtained at the output of the Transformer encoder ($\mathbf{z}_L^0$) (Eq. 4). For classification, an MLP with one hidden layer during pre-training, replaced by a single linear layer in fine-tuning, is applied to $\mathbf{z}_L^0$.

Positional embeddings are added to each patch embedding to retain spatial information, forming the complete input sequence for the encoder.

The Transformer encoder consists of alternating layers of multi-head self-attention (MSA) and MLP blocks (Eq. 2, 3), each preceded by Layer Normalization (LN) and followed by a residual connection. The MLP block has two layers and uses a GELU activation function.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{1}$$

$$\mathbf{z'}_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1\ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \qquad \ell = 1\ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

Inductive Bias. The ViT model incorporates less image-specific inductive bias compared to CNNs. In CNNs, characteristics such as locality, two-dimensional neighborhood structure, and translation equivariance are explicitly integrated across layers, shaping the model's entire structure. In contrast, ViT models have only local and translationally equivariant biases within the MLP layers, while the self-attention layers operate on a global scale.

Hybrid Architecture. An alternative to using raw image patches as input involves employing feature maps derived from a CNN. In this hybrid approach, the patch embedding projection $\mathbf{E}$ (Eq. 1) is applied to patches that are extracted from a CNN feature map, combining the benefits of CNNs and Transformers.

4. Experiments

4.1. Comparison to State of The Art

	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: Performance comparison of Vision Transformer and ResNet-based models on popular benchmarks. Vision Transformer pre-trained on JFT-300M achieves superior results while requiring significantly less compute, and performs competitively when pre-trained on ImageNet-21k.

4.2. Scaling Study

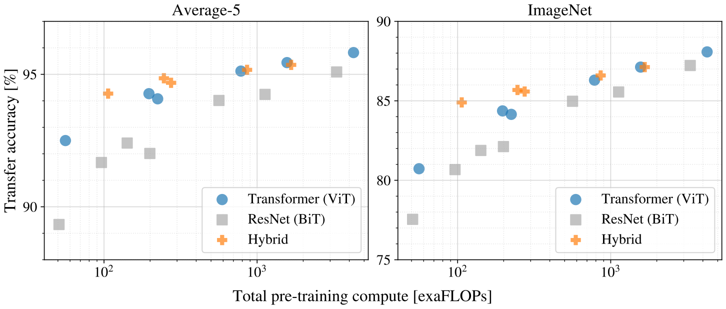

Figure 1: Performance vs. compute for Vision Transformer, ResNets, and hybrid models. Vision Transformer consistently outperforms ResNets within the same compute budget, with hybrids benefiting smaller models.

4.3. Inspecting Vision Transformer

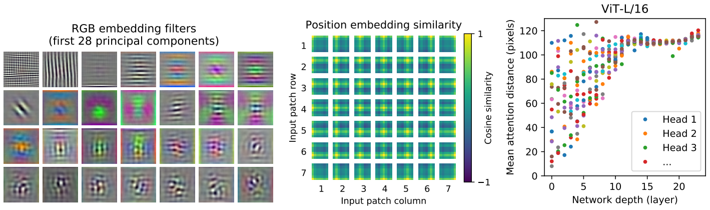

Figure 2: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer.

The first layer of the ViT linearly projects the flattened patches into a lower-dimensional space (Eq. 1). Figure 2 (left) shows the top principal components of the the learned embedding filters.

After the projection, a learned position embedding is added to the patch representations. Figure 2 (center) shows that the model learns to encode distance within the image in the similarity of position embeddings. Patches in the same row/column have similar embeddings.

The mean attention distance is calculated to analyze how self-attention allows ViT to combine information across the entire image (Figure 2, right). This "attention distance" is an analogue to the receptive field size in CNNs. Even in the lower layers, some attention heads exhibit broad coverage, indicating ViT's capability for global spatial integration.

[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.