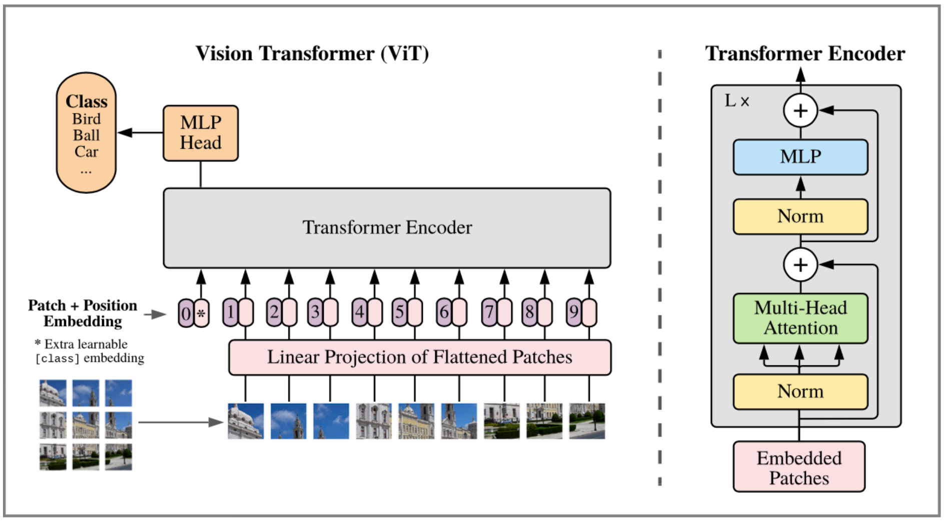
An Image is worth 16×16 Words: Transformers for Image Recognition at Scale
(Vision Transformer)

1. Model



ViT Overview

1. Patch Embedding (2D image —> 1D sequence)

- $H \times W \times C \rightarrow N \times (P^2 \times C)$ $\quad$ ($N = \frac{HW}{P^2}$; number of patches)

- <u>Can use feature maps of a CNN as an alternative to raw image patches</u>

- Linear projection to the vector size D (since every layer uses vector size D.)

- Prepend extra learnable [class] embedding

- Add 1D position embeddings

- $Z_0 = [X_{class}; X^1E; X^2E; \ldots; X^NE] + E_{pos}$

2. Multi-headed Self-Attention and Multi-layer Perceptron

- $Z_l' = MSA(LN(Z_{l-1})) + Z_{l-1}$

- $Z_l = MLP(LN(Z_l')) + Z_l'$

- Note the residual connections after every block

3. Output: $y = LN(Z_L^0)$

4. Fine Tuning

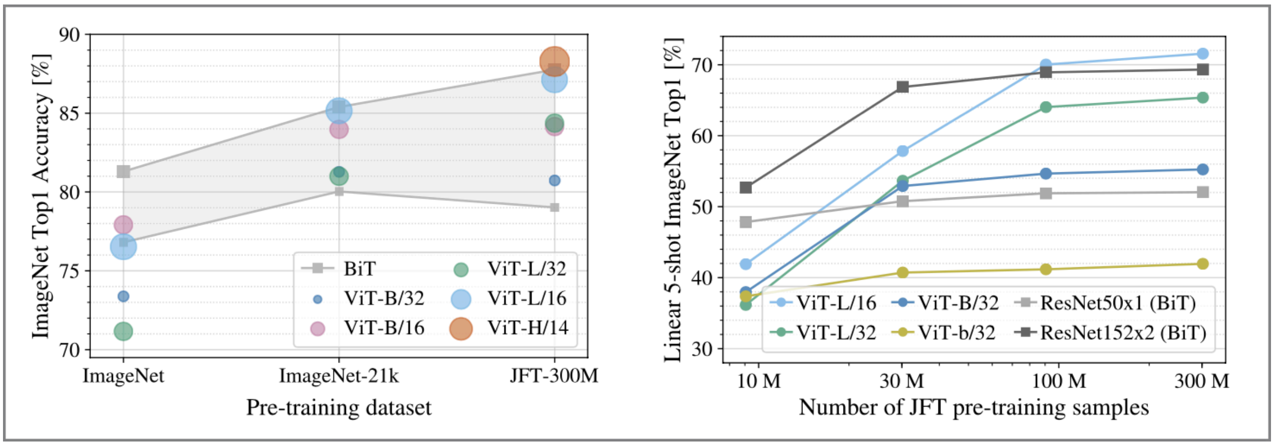
- Remove the pre-trained prediction head and attach a zero-initialized $D \times K$ FC layer (—> Softmax)

- Higher resolution —> Longer sequence length
  - pre-trained position embeddings can be useless —> 2D interpolation

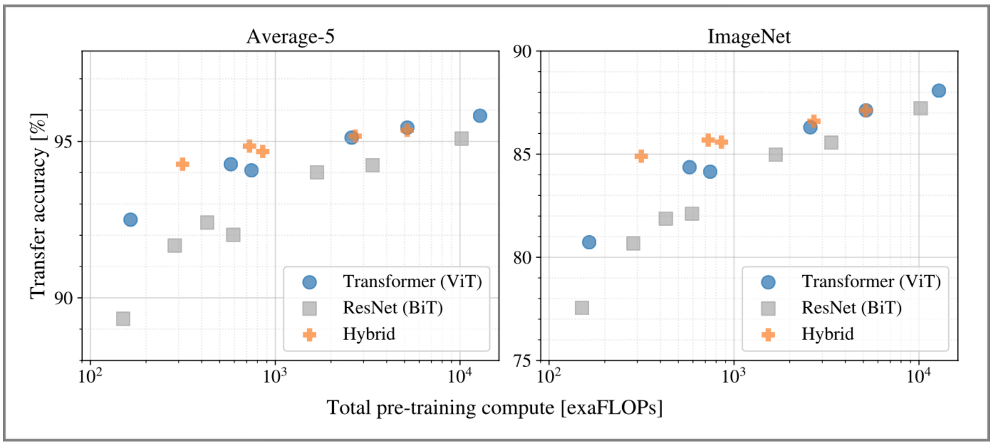
2. Evaluation

	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** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Comparison with SOTA



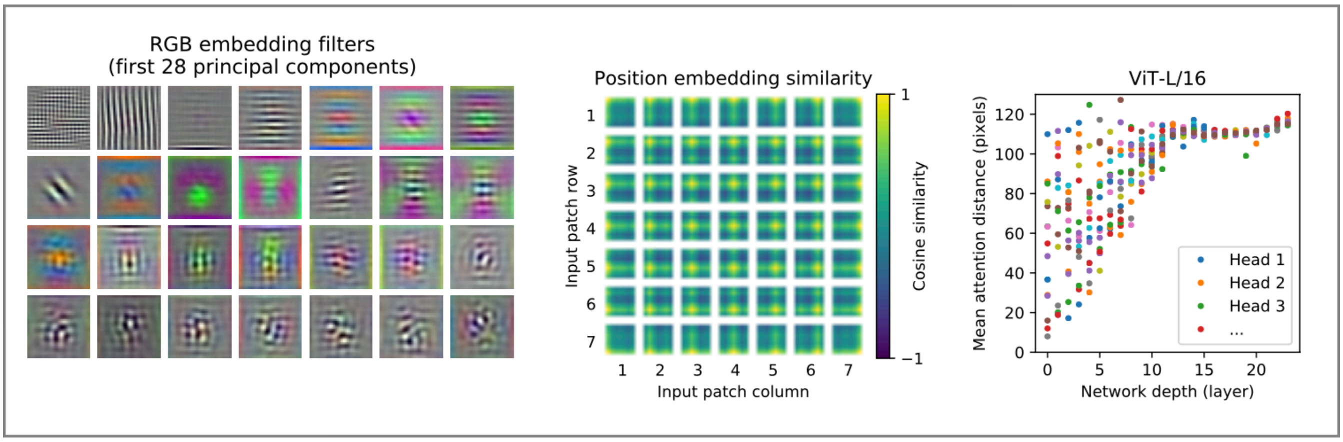
Pre-training size and Accuracy



Scaling study

- Convolutional inductive bias is useful for smaller datasets, but for larger ones, learning the relevant patterns is sufficient
- For small computational budgets, hybrid is best, but the difference vanishes for larger models.
- ViT is not saturated. Can be better scaled.

3. Further Studies



Inspecting ViT

1. Convolution layer 없이 linear projection 만으로도 가로선, 세로선 등을 학습하는 것을 보아 CNN의 low layer와 비슷한 역할을 수행함을 알 수 있다.
2. Position embedding similiarity for same rows and columns is high.
3. Self-attention을 통해 이미지의 전역적인 특징을 추출할 수 있는가? Both highly localized and globally integrated heads are discovered even in the lowest layers.
4. Hybrid model (+ResNet) : 전역추출 효과적 (residual blocks)