An Image is Worth 16x16 Words: Transformers For Image Recognition At Scale

Anonymous authors

Song Wang, Liyan Tang

https://openreview.net/pdf?id=YicbFdNTTy

Reliance on CNNs is not necessary and a pure transformer can perform well on image classification tasks when applied directly to sequences of image patches.

Background

Self-attention based architectures have become model of choice in NLP.

Pre-train on a large text corpus, and fine-tune on a smaller task-specific dataset.

In Computer Vision, convolutional architectures remain dominant.

Method - Vision Transformer

Split an image into patches

Input sequence of linear embeddings of patches to a Transformer

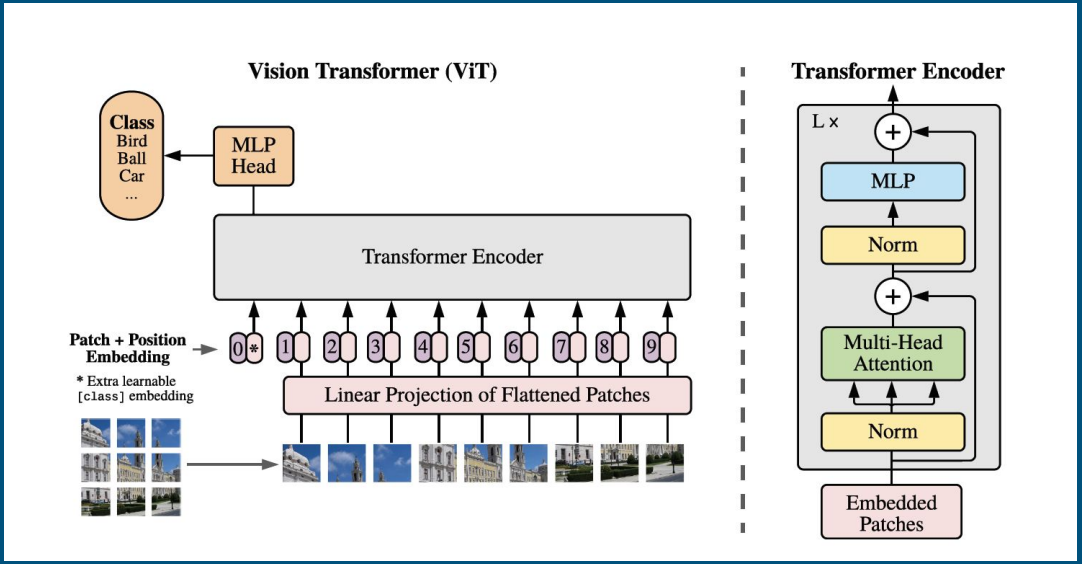
Image patches are treated the same way as tokens(words) in NLP

Train the model in supervised fashion

Limits: Do not generalize well when trained on insufficient amounts of data.

Method - Vision Transformer

Method - Vision Transformer



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

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

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

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

Method - Hybrid Architecture

One of the intermediate 2D feature maps of the ResNet is flattened into a sequence

Projected to the Transformer dimension

Fed as an input sequence to a Transformer

Method - Fine-tuning

Pre-train on large datasets, fine-tune to smaller downstream tasks

Remove pre-trained prediction head

Attach a zero-initialized DxK feedforward layer, K is the number of downstream classes

Method - Higher resolution

Keep patch size the same

A larger effective sequence length -> pre-trained position embeddings may no longer be meaningful

2D interpolation of the pre-trained position embeddings, according to their locations in the original image

Experiments Set Up

16x16 input patch size

Model	Layers	Hidden size 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: Configuration of our different model variants.

BiT-L: Big Transfer ResNet (SOTA CNNs from the literature)

Comparison to SOTA

	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	$88.4/\mathbf{88.5}^*$
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	$\mathbf{99.50} \pm 0.06$	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	$\mathbf{94.55} \pm 0.04$	93.90 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	$\mathbf{97.56} \pm 0.03$	97.32 ± 0.11	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	$\mathbf{99.74} \pm 0.00$	99.63 ± 0.03	–
VTAB (19 tasks)	$\mathbf{77.16} \pm 0.29$	75.91 ± 0.18	76.29 ± 1.70	–
TPUv3-days	2.5k	0.68k	9.9k	12.3k

ViT-L/16: ViT-L with 16x16 patch size.

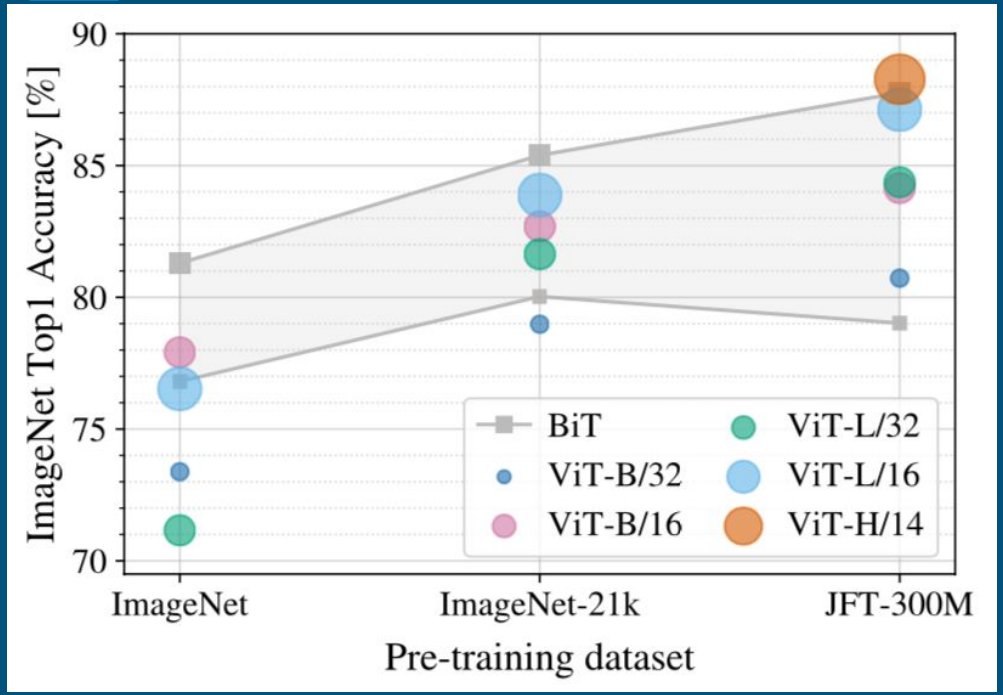
TPUv3-days: the number of TPUv3-days to pre-train each model.

Comparison to SOTA

1. The smaller ViT-L/16 model matches or outperforms BiT-L on all datasets, while requiring substantially less computational resources to train.

2. The larger model, ViT-H/14, further improves the performance (because of larger model and smaller input patch size).

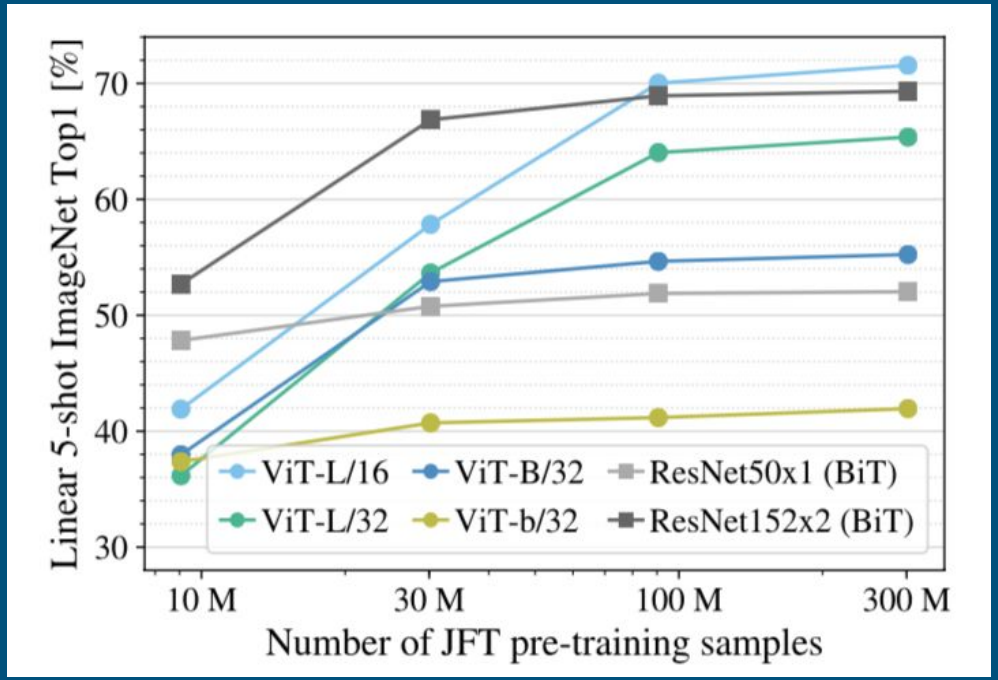
Pre-training Data Requirements



Shades Area: the performance region spanned by BiT models of different sizes.

1. When pre-trained on the smallest dataset, ImageNet, ViT-Large models underperform compared to ViT-Base models.
2. With JFT-300M, ViT finally overtakes BiT.

Pre-training Data Requirements

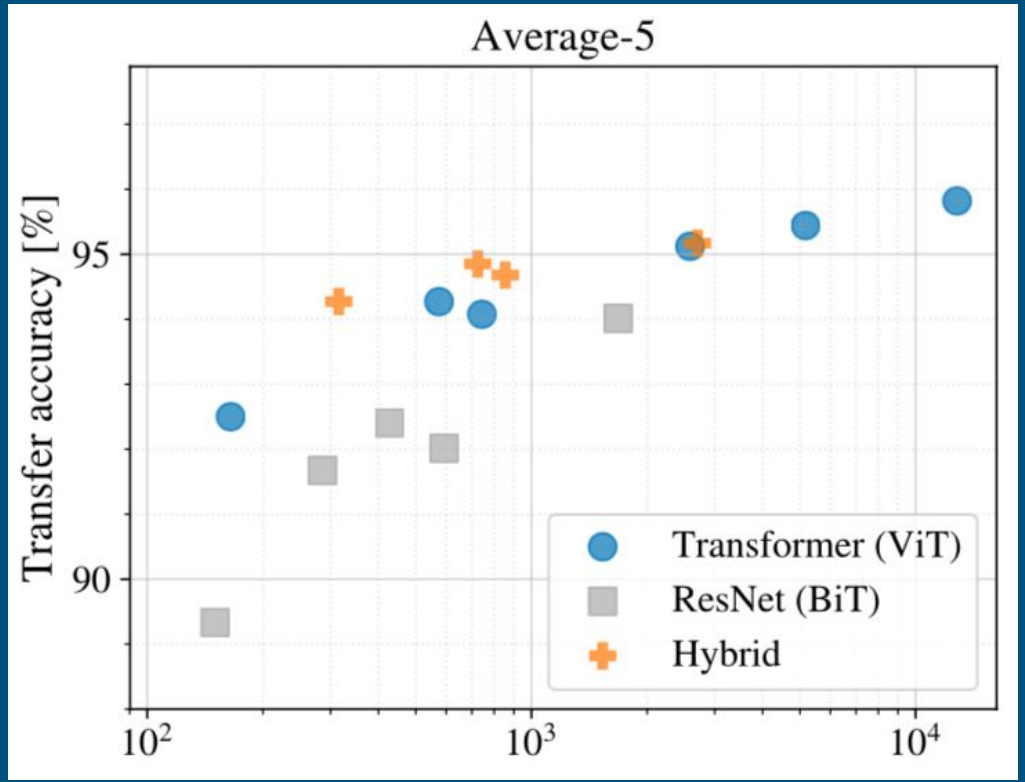


Linear few-shot evaluation on ImageNet versus pre-training size:

ResNets perform better with smaller pre-training datasets but plateau sooner than ViT which performs better with larger pre-training.

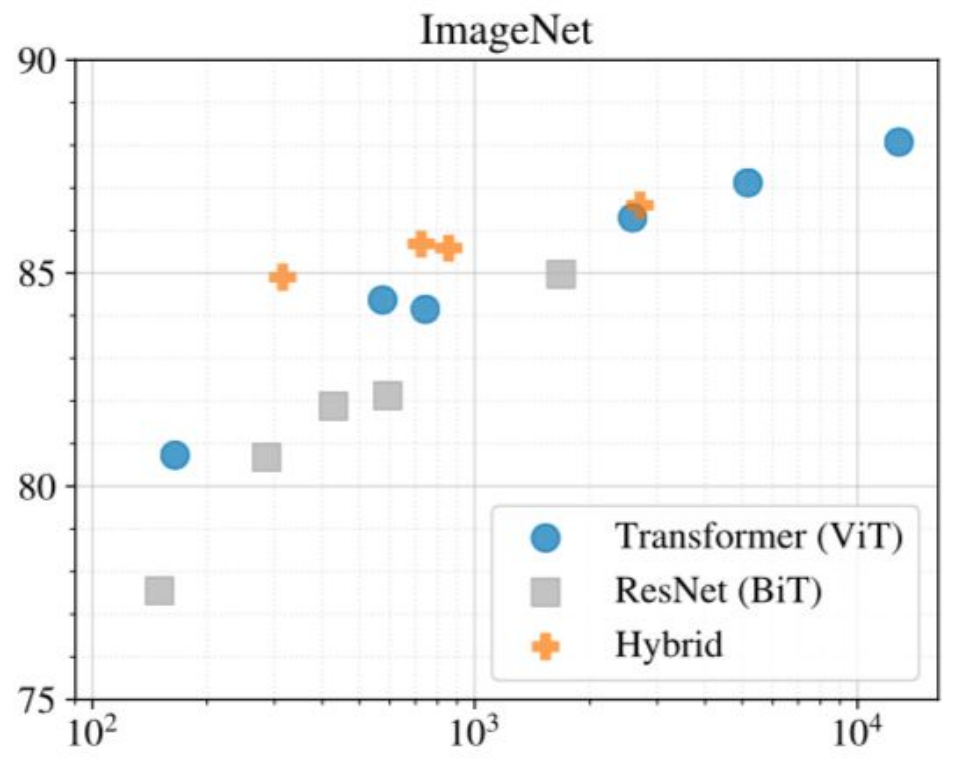
This result reinforces the intuition that the convolutional inductive bias is useful for smaller datasets, but for larger ones, learning the relevant patterns is sufficient, even beneficial.

Scaling Study



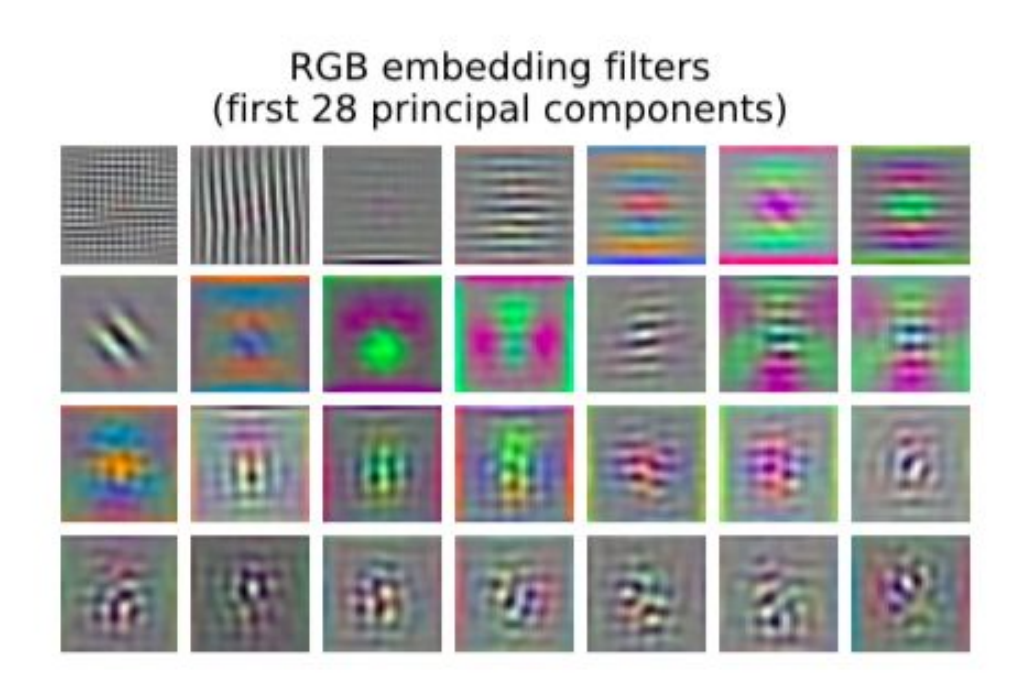
1. Vision Transformer dominate ResNets on the performance/compute trade-off. ViT uses approximately 2× less compute to attain the same performance.
2. Hybrids slightly outperform ViT at small computational budgets, but the difference vanishes for larger ones (local feature/attention/bias would be help helpful as more data being trained on).
3. ViT performance does not seem to be saturating with the increased model size.

Scaling Study



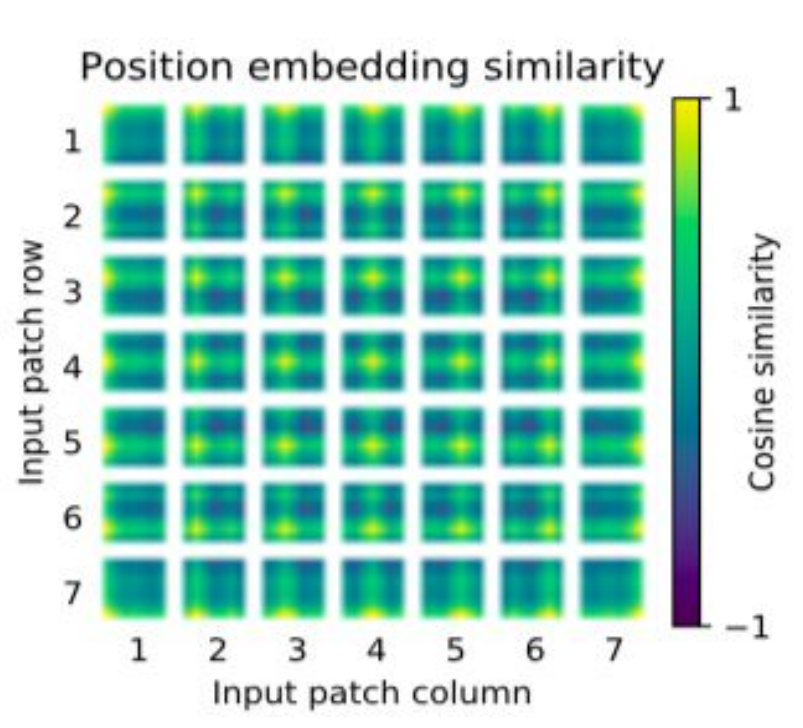
1. Vision Transformer dominate ResNets on the performance/compute trade-off. ViT uses approximately 2× less compute to attain the same performance.
2. Hybrids slightly outperform ViT at small computational budgets, but the difference vanishes for larger ones (local feature/attention/bias would be help helpful as more data being trained on).
3. ViT performance does not seem to be saturating with the increased model size.

Inspect Vision Transformer



Similar to CNN, ViT could learn basic functions of the image structures within each patch.

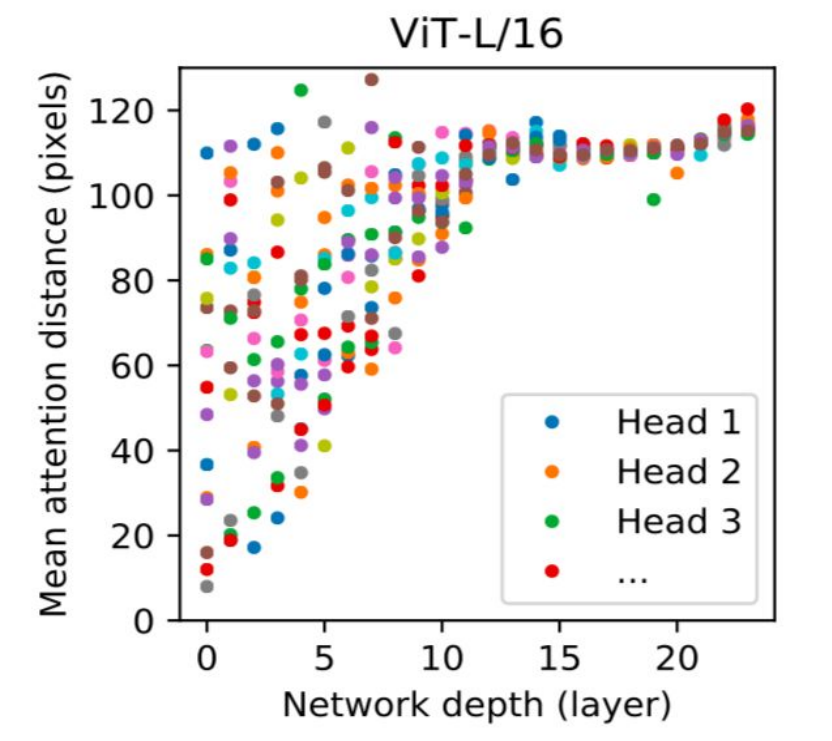
Inspect Vision Transformer



The model learns to encode distance within the image in the similarity of position embeddings. That is,

1. closer patches tend to have more similar position embeddings.
2. The row-column structure appears; patches in the same row/column have similar embeddings.

Inspect Vision Transformer



Compute the average distance based on the attention weights.

Some heads attend to most of the image already in the lowest layers, showing that the ability to integrate information globally is indeed used by the model.

Conclusions

1. Explored the direct application of Transformers to image recognition, where they do not introduce any image-specific inductive biases into the model architecture.
2. Interpret an image as a sequence of patches and process it by a standard Transformer encoder as used in NLP.
3. Vision Transformer matches or exceeds the state of the art on many image classification datasets, whilst being relatively cheap to pre-train.

Future Works

1. Apply Visual Transformer to other computer vision tasks, such as detection and segmentation.
2. Continue exploring self-supervised pre-training methods. There is still large gap between self-supervised (*e.x. Masked patch prediction*) and large-scale supervised pre-training.
3. To further scale ViT, given that the performance does not seem yet to be saturating with the increased model size.

Questions?