AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ETRI, Artificial Intelligence, Seungeun Han

과학기술연합대학원대학교 UNIVERSITY OF SCIENCE & TECHNOLOGY

한국전자통신연구원 Electronics and Telecommunications Research Institute

Contents

과학기술연합대학원대학교
UNIVERSITY OF SCIENCE & TECHNOLOGY

한국전자통신연구원
Electronics and Telecommunications Research Institute

1. Introduction

- "**Transformers**" become the state of the art method in many NLP (Natural Language Processing) tasks.

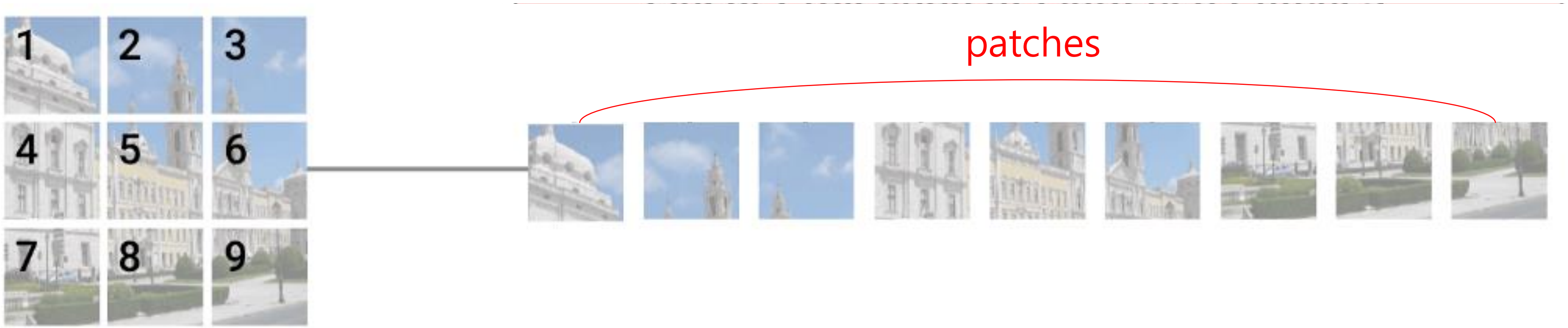
Attention is **all you need**
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …
☆ 저장 99 인용 57192회 인용 관련 학술자료 전체 46개의 버전

- However, In computer vision, Convolutional architectures remain dominant. (CNN)

- So, the idea is that if we apply a standard Transformer directly to images with the fewest modifications, it will be better than CNN.

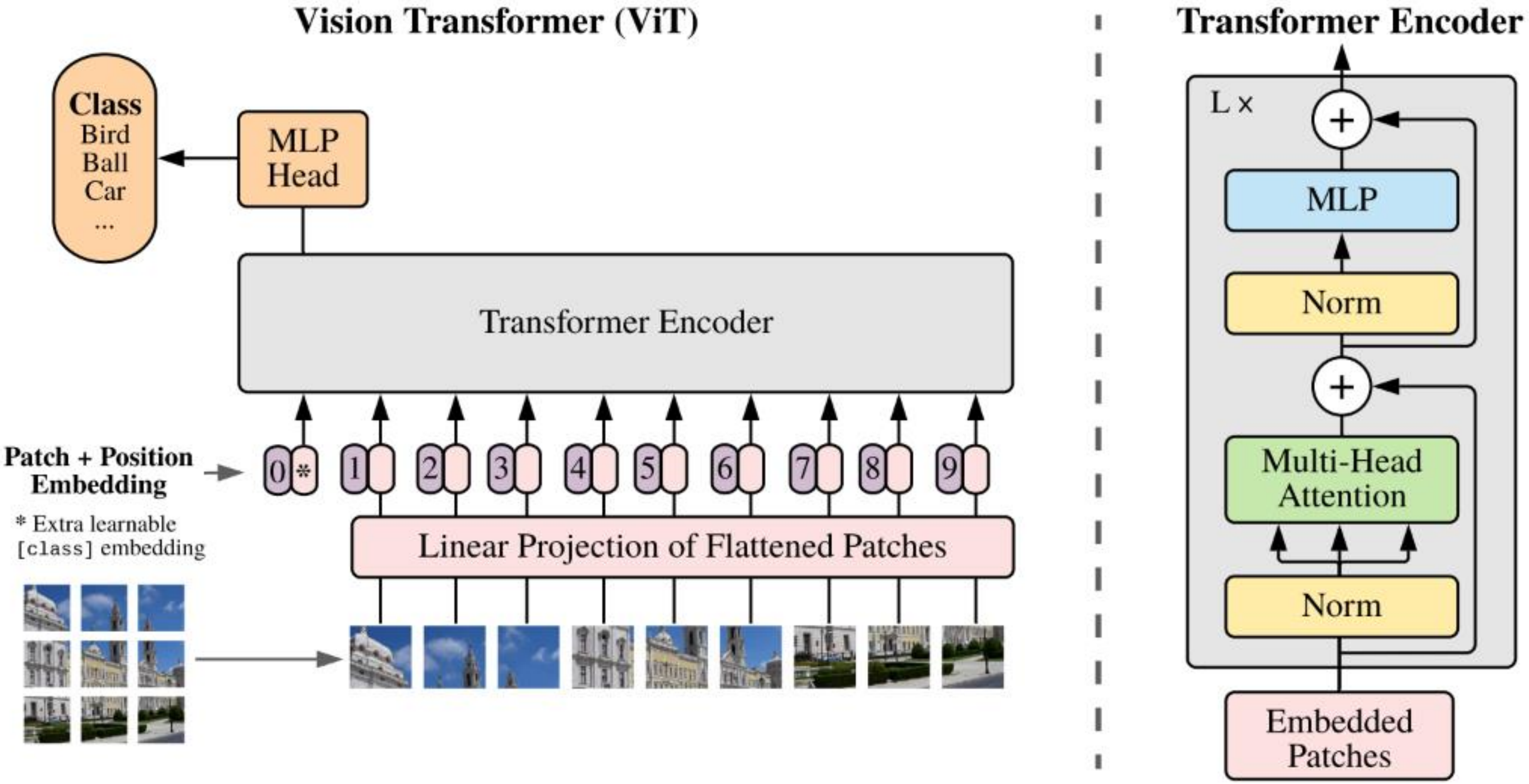
1. Introduction

- Split an image into patches.
- The patches ≈ tokens (words) in an NLP



- Provide the sequence of linear embeddings of these patches.

2. Vision Transformer(ViT)



2. Vision Transformer(ViT)

2. Vision Transformer(ViT)

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

$$\mathbf{z}'_\ell = \underbrace{\mathrm{MSA}}_{\text{Multi-head Attention}}(\underbrace{\mathrm{LN}}_{\text{Layer Normalization}}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1\ldots L \tag{2}$$

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

$$\underbrace{\mathbf{y}}_{\text{Predicted class}} = \mathrm{LN}(\mathbf{z}_L^0) \tag{4}$$

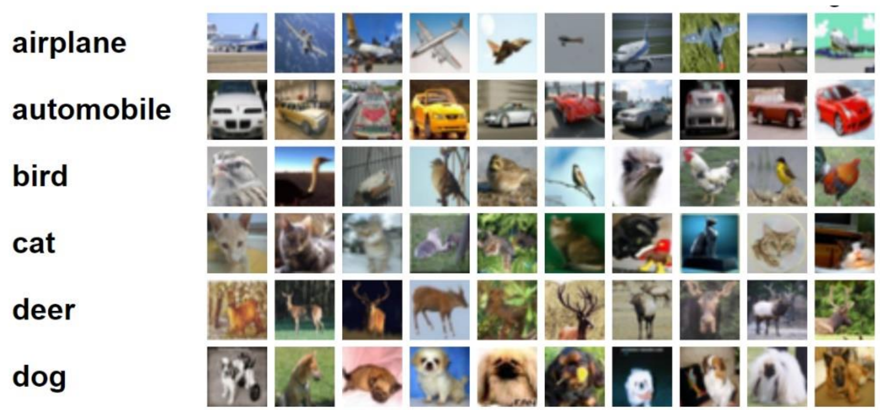
3. Experiments

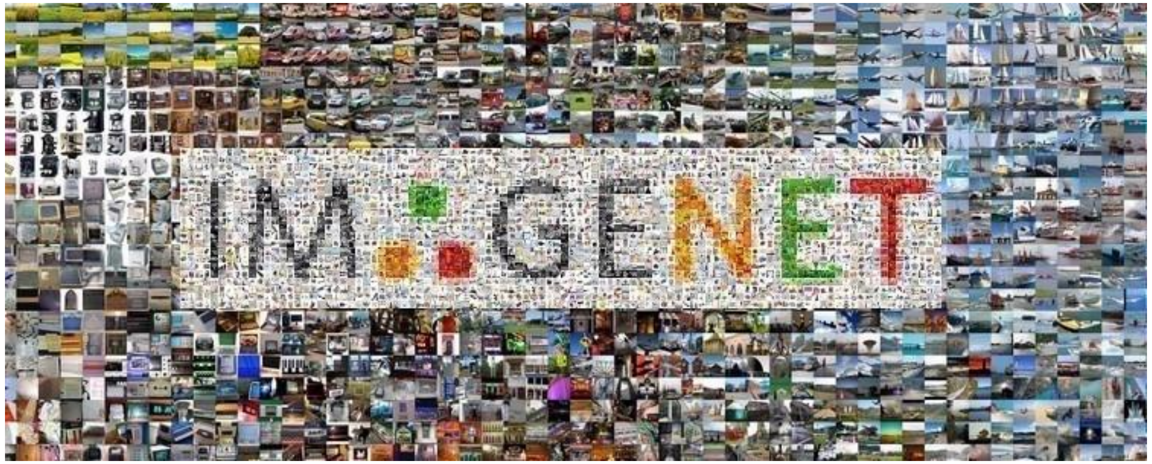
▪ Datasets

- ILSVRC-2012 ImageNet dataset

  : 1k classes and 1.3M images
- JFT

  : 18k classes and 303M high-resolution images
- ImageNet
- CIFAR-10/100
- Oxford-IIIT Pets
- Oxford Flowers-102





3. Experiments

- ResNet(CNN) VS ViT VS Hybrid

- 19-task VTAB classification

- Model Variants

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: Details of Vision Transformer model variants.

- ViT-L/16 : the "Large" variant with 16x16 input patch size

3. Experiments

- Metrics

- Fine-tuning Accuracy

: captures the performance of each model after fine-tuning it on the respective dataset.

- Few-shot Accuracy

: obtained by solving a regularized least-squares regression problem.

- Mainly using fine-tuning accuracy.
- But, sometimes using few-shot accuracy for fast evaluation.

4. Results

- Comparison to state of the art

Pre-trained dataset

	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

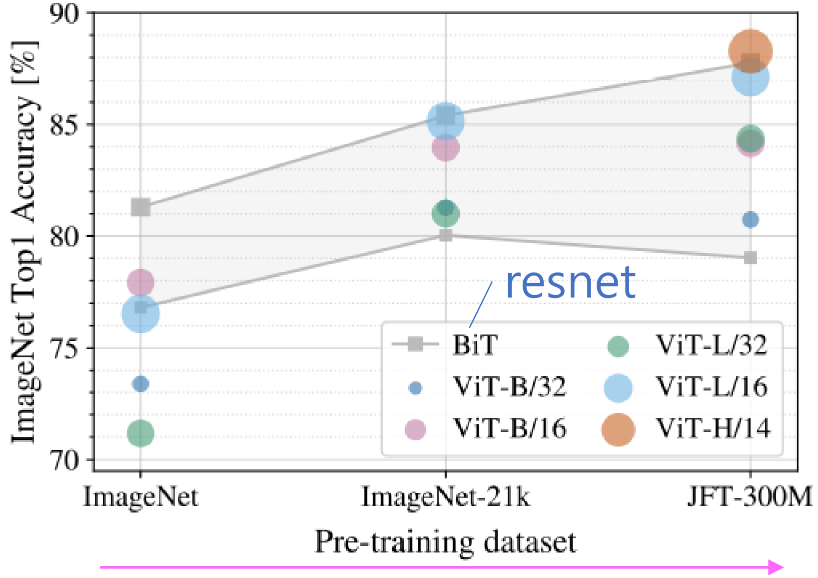
By this, we can see computational burden.

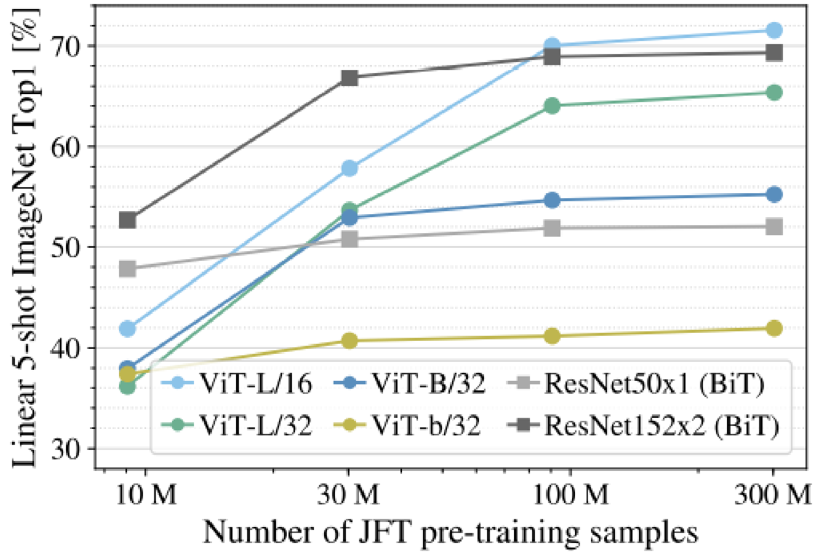
- ViT-H/14 improves the performance on ImageNet, CIFAR-100, and the VTAB suite.
- ViT-L/16 model pre-trained on the public ImageNet-21k dataset performs well on most datasets too, while taking fewer resources to pre-train.

4. Results

- how crucial is the dataset size?

Accuracy increase



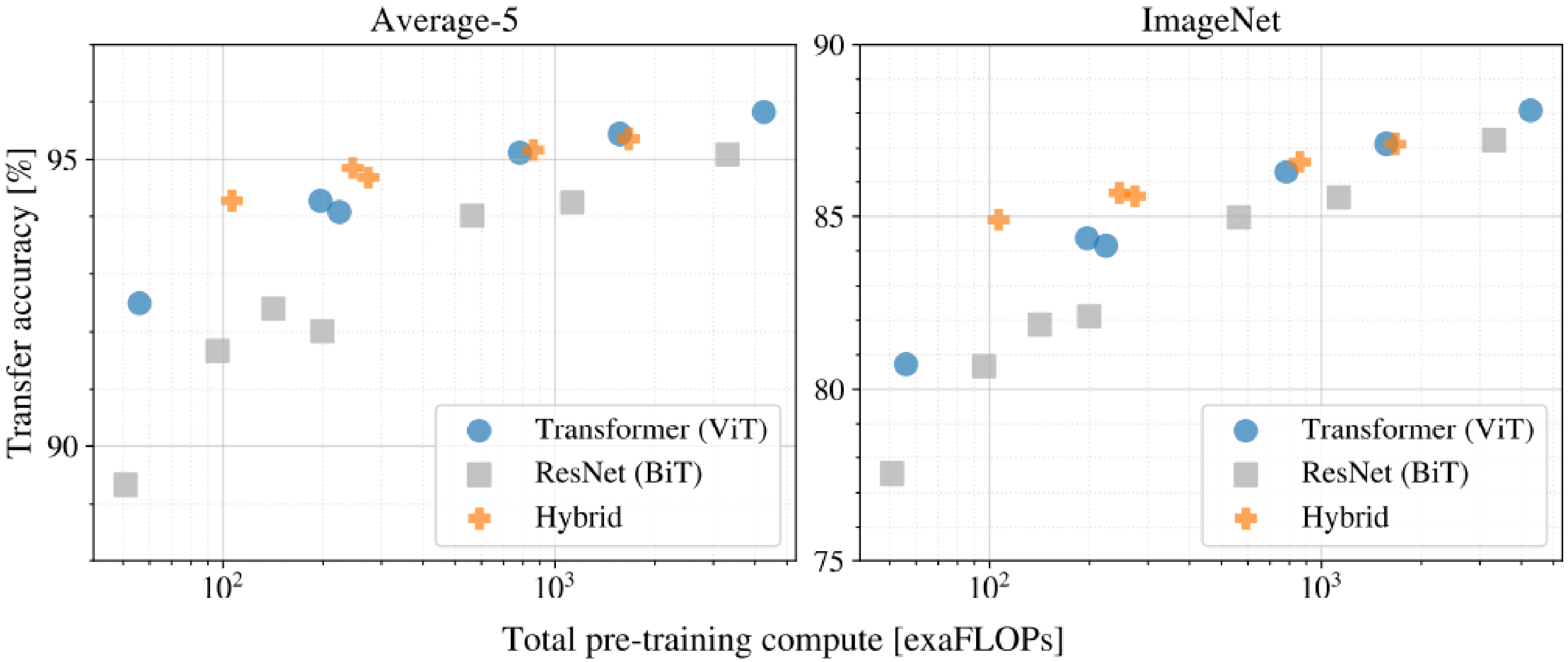


Size of dataset increase

- The CNN outperforms ViT on ImageNet, but with the larger datasets, ViT overtakes.

4. Results

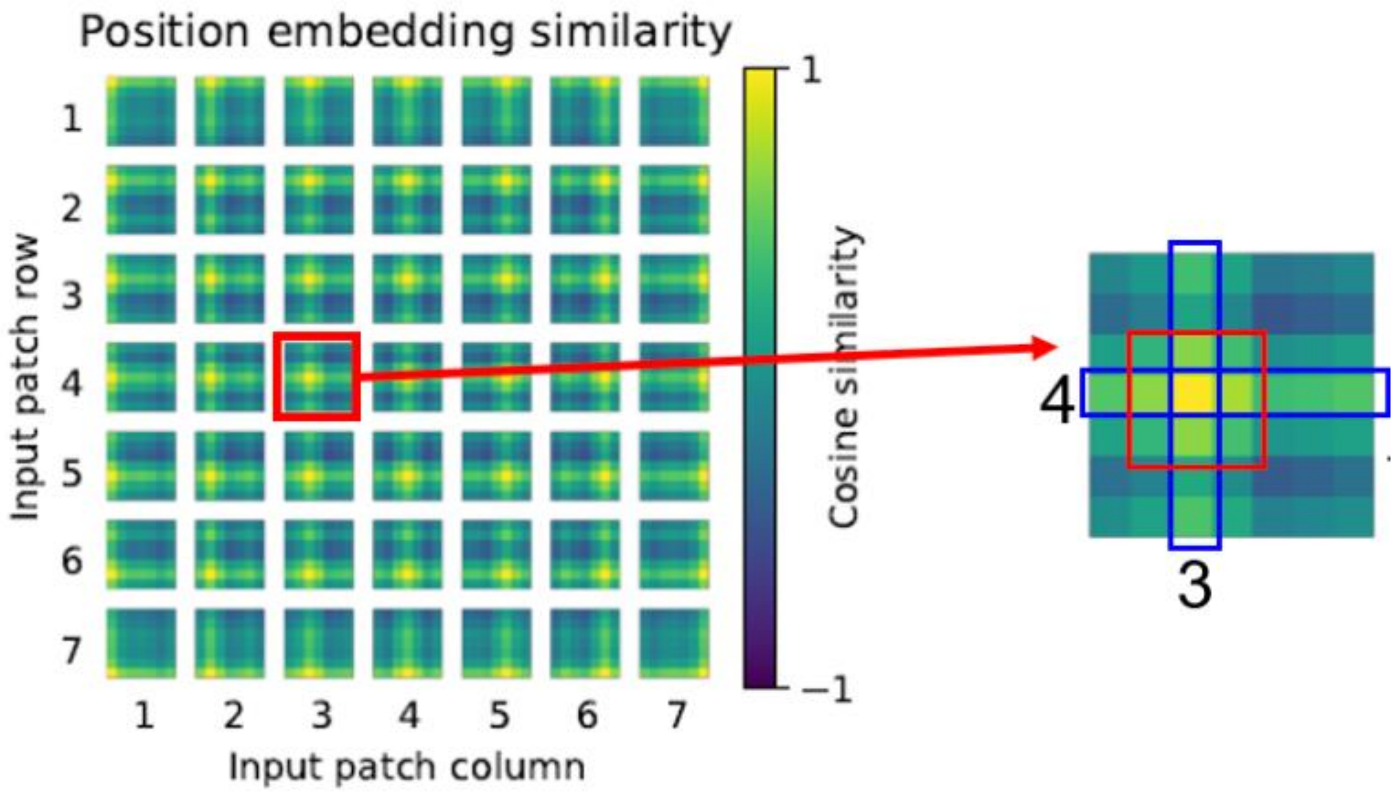
- Scaling Study



1. Vision Transformers dominate ResNets on the performance/compute trade-off.
2. Hybrids slightly outperform ViT at small, but the difference vanishes for larger models.
3. Vision Transformers appear not to saturate and can be scaled.

4. Results

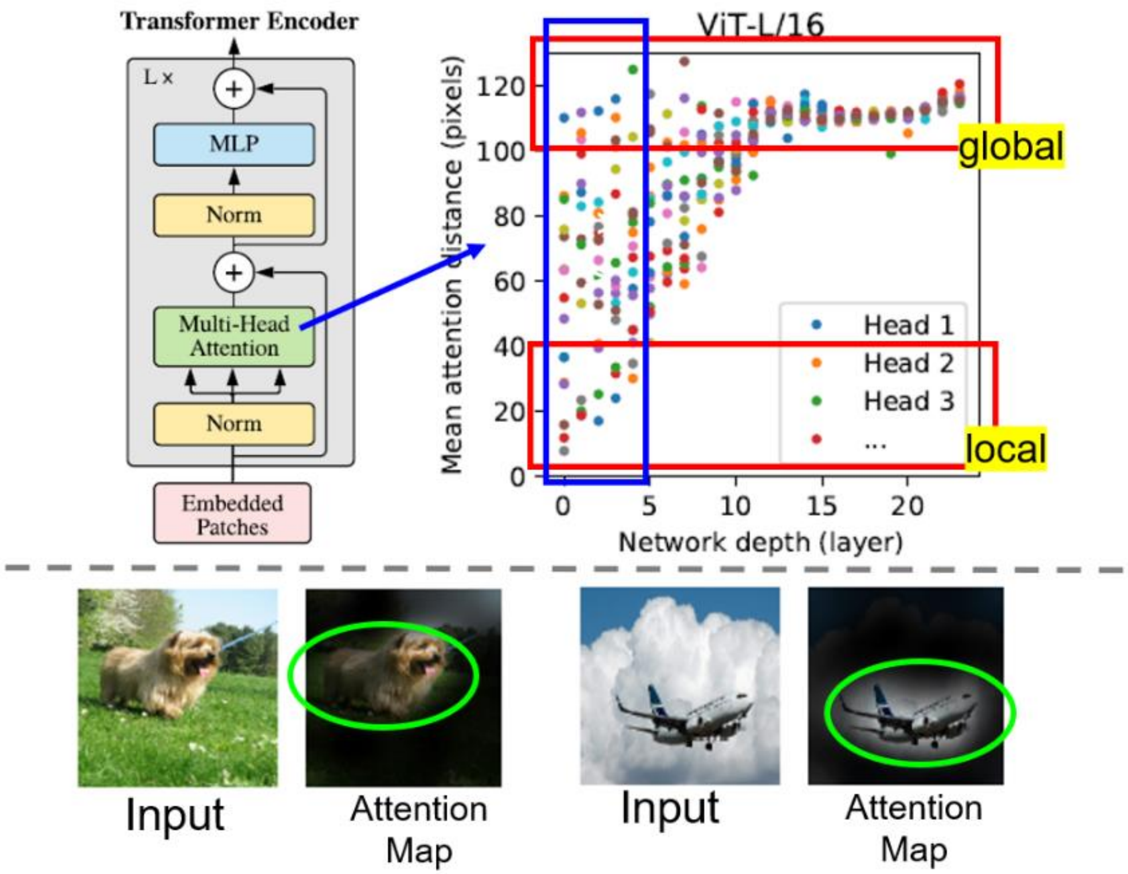
- How the Vision Transformer processes image data
  - Similarity of position embeddings



1. Closer patches tend to have more similar position embeddings.
2. Same row/column patches have similar embeddings.
3. The 1D position embeddings learn to represent 2D image topology

4. Results

- How the Vision Transformer processes image data
  - Mean attention distance



- Attention distance ≈ receptive field size in CNN
- Self-attention allows ViT to integrate information across the entire image even in the lowest layers.

5. Conclusion

- On large dataset, ViT is better than state-of-the-art CNN.
- While requiring lower computational resources.
- Simple, Scalable
- Thus, Vision Transformer exceeds the state of the art on many image classification datasets, being relatively cheap to pre-train.

Thank you