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*,†
Google Research

DSLAB 이수아

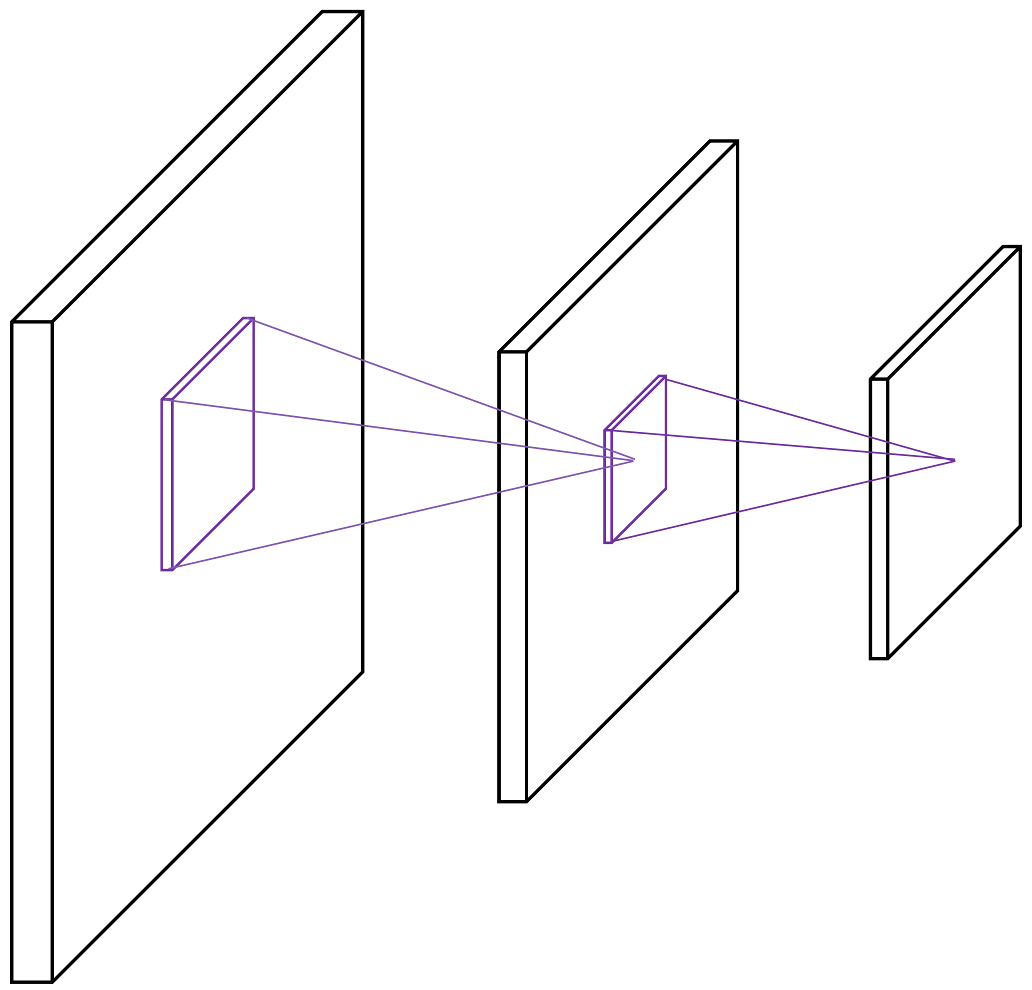
Abstract

Vision Transformer (ViT)

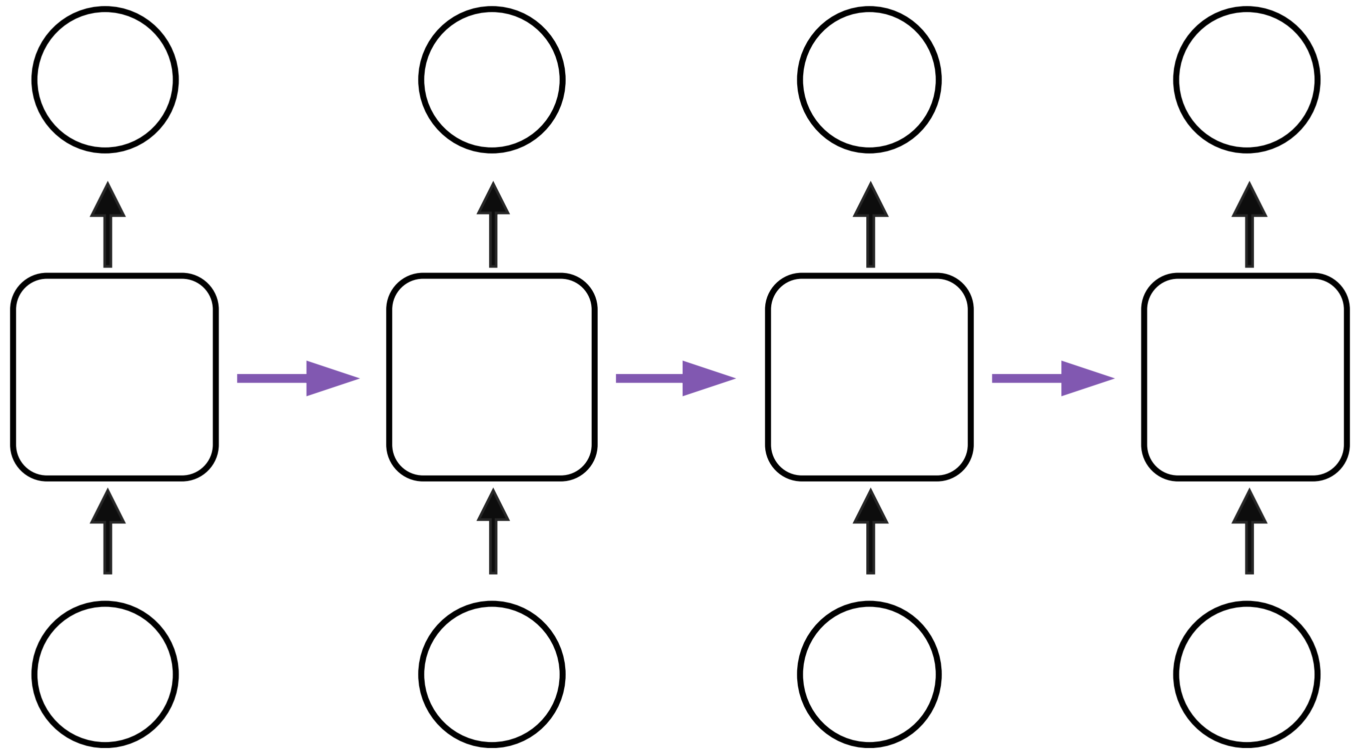
- Transformer 구조 활용
- SOTA의 CNNs보다 비슷하거나 뛰어난 성능
- Large-scale(14M~300M) → mid or small sized(ImageNet, CIFAR-100)

Inductive Bias

- **Inductive bias :** 새로운 데이터에 대해 좋은 성능을 내기 위해 모델에 사전적으로 주어지는 가정
  - CNN : Locality

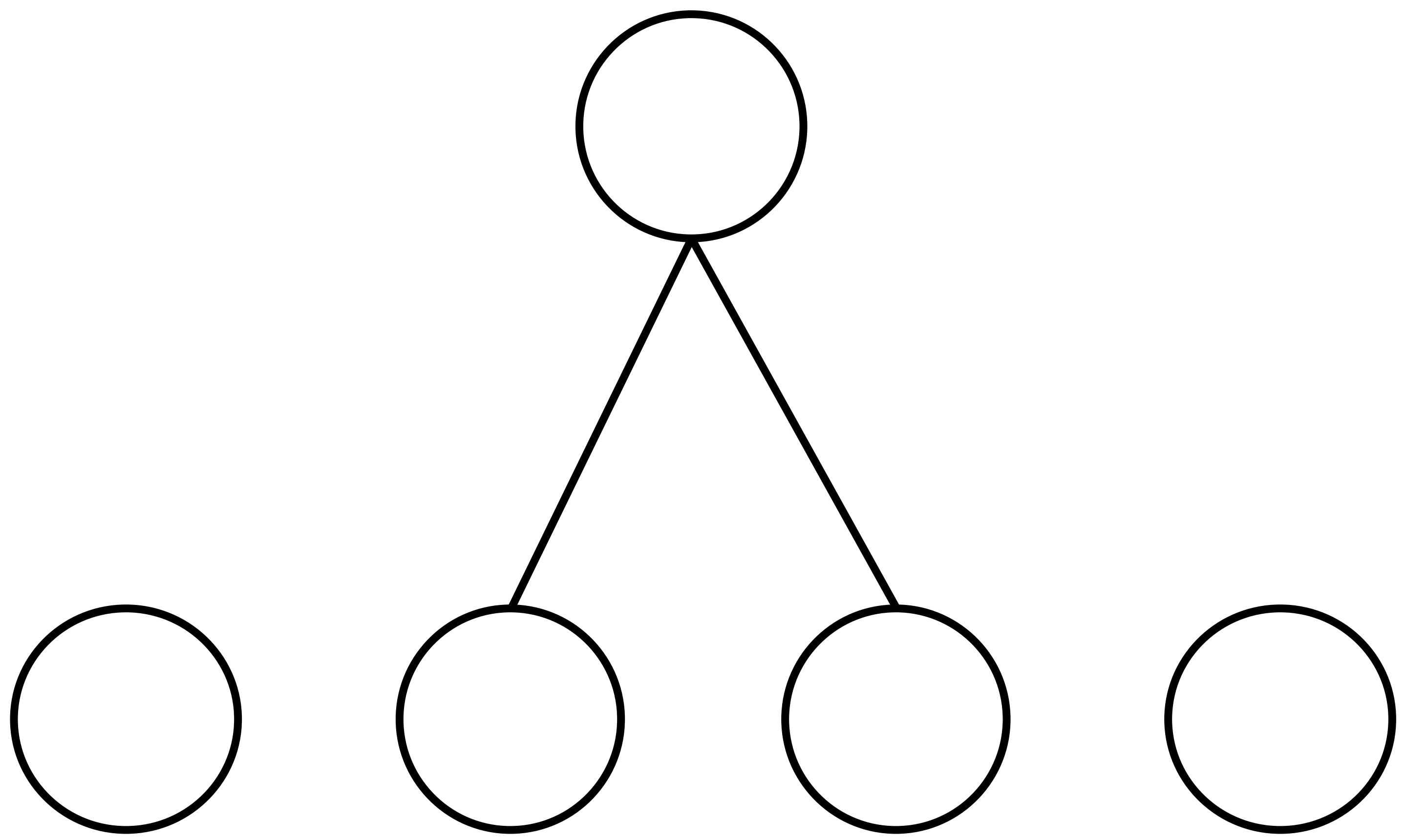


  - RNN : Sequentiality

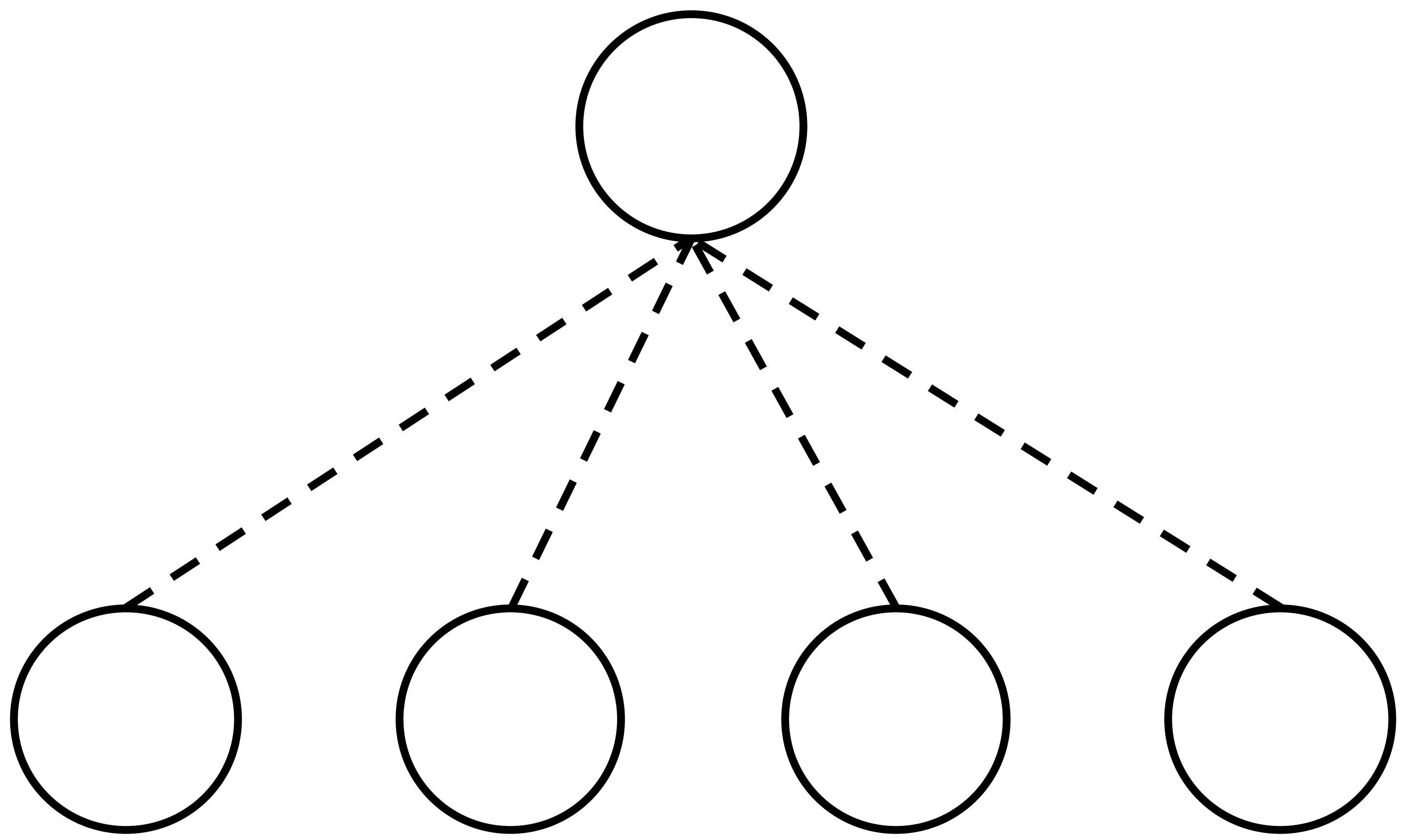


Inductive Bias

- CNN : Locality ➡ Global ↓
- Transformer : self-attention ➡ inductive bias ↓



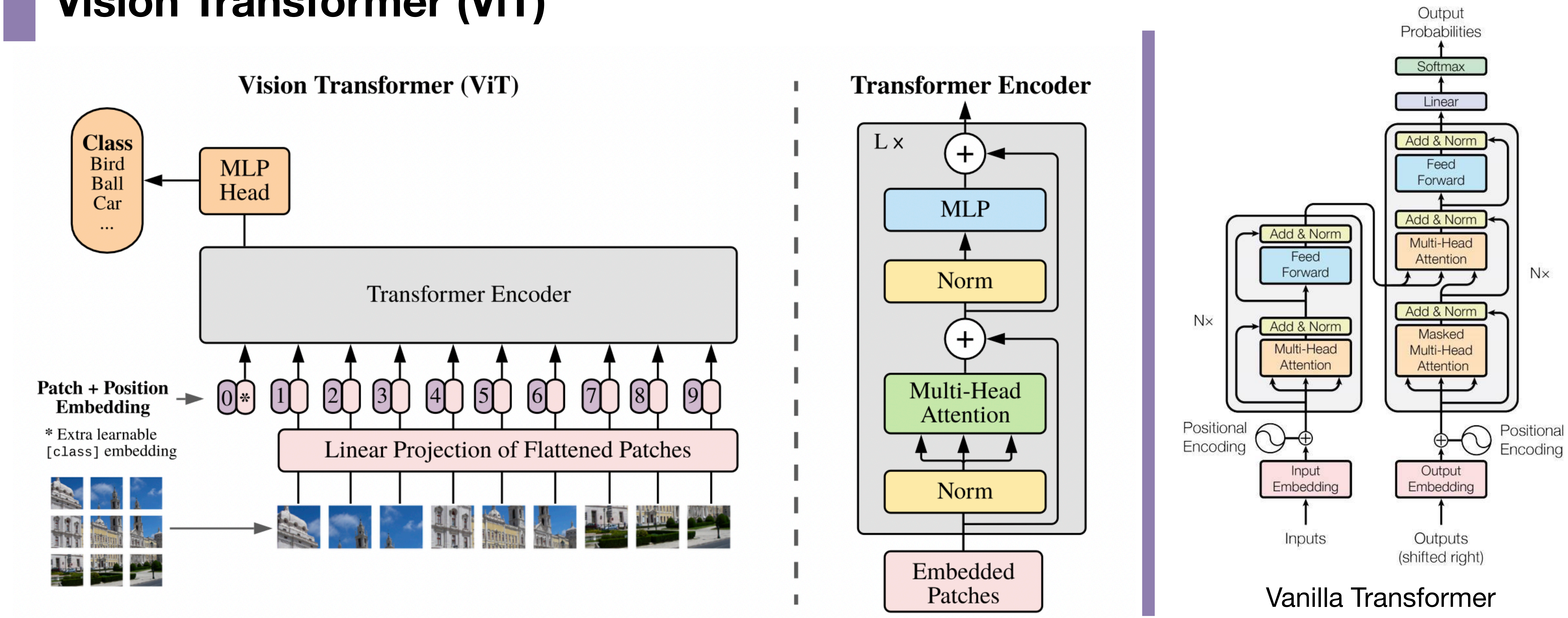
CNN



Transformer

Method

Vision Transformer (ViT)



Vanilla Transformer

Method

Vision Transformer (ViT)

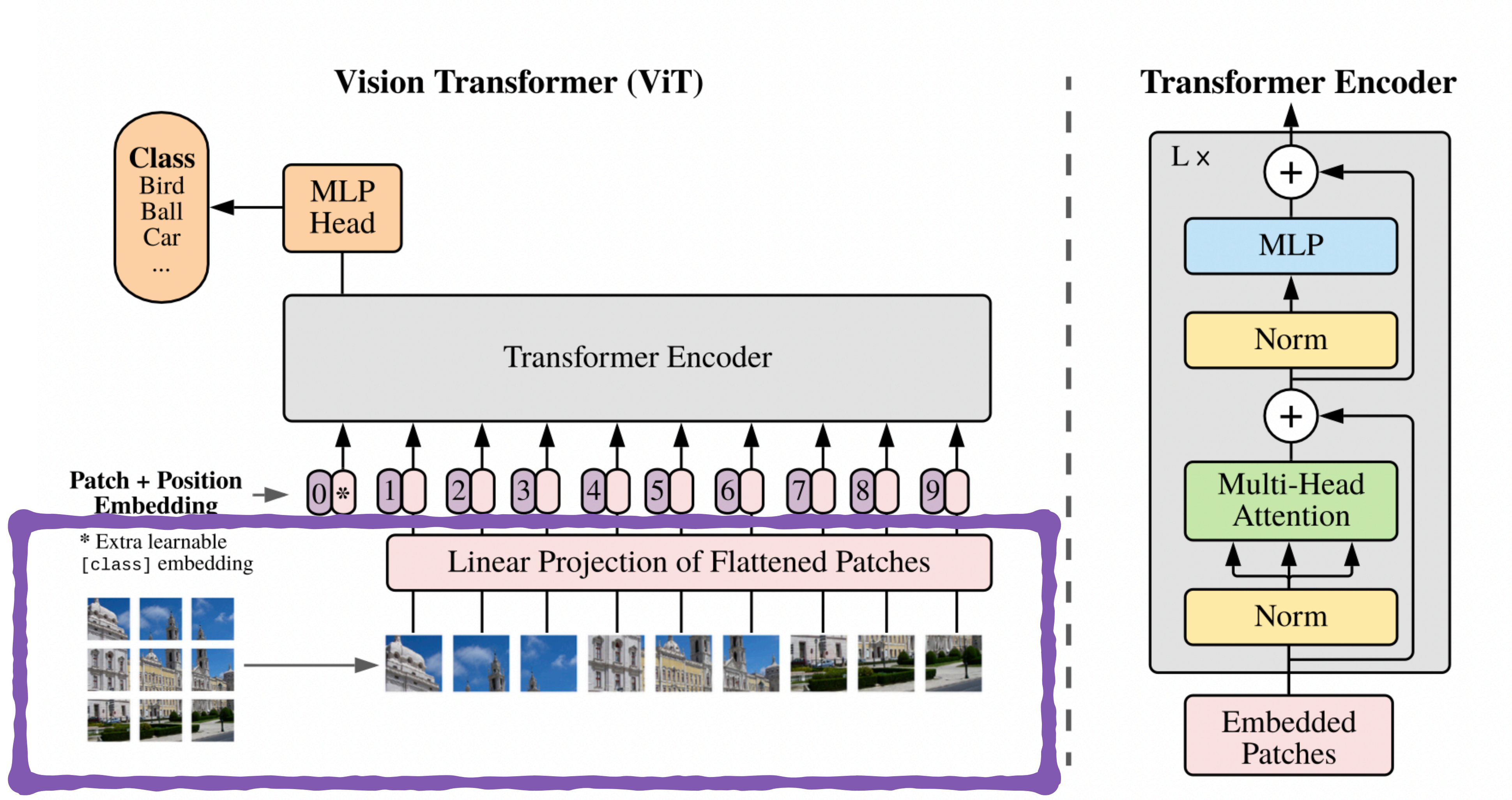


image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

▶ flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$

(H, W) : resolution of the original image
(P, P) : resolution of each image patch
$N = HW/P^2$: resulting number of patches

latent vector size D

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}, \quad E \in R^{(P^2 \cdot C) \times D}, E_{pos} \in R^{(N+1) \times D}$$

$x_p^1 \in R^{1 \times (P^2 \times C)}$

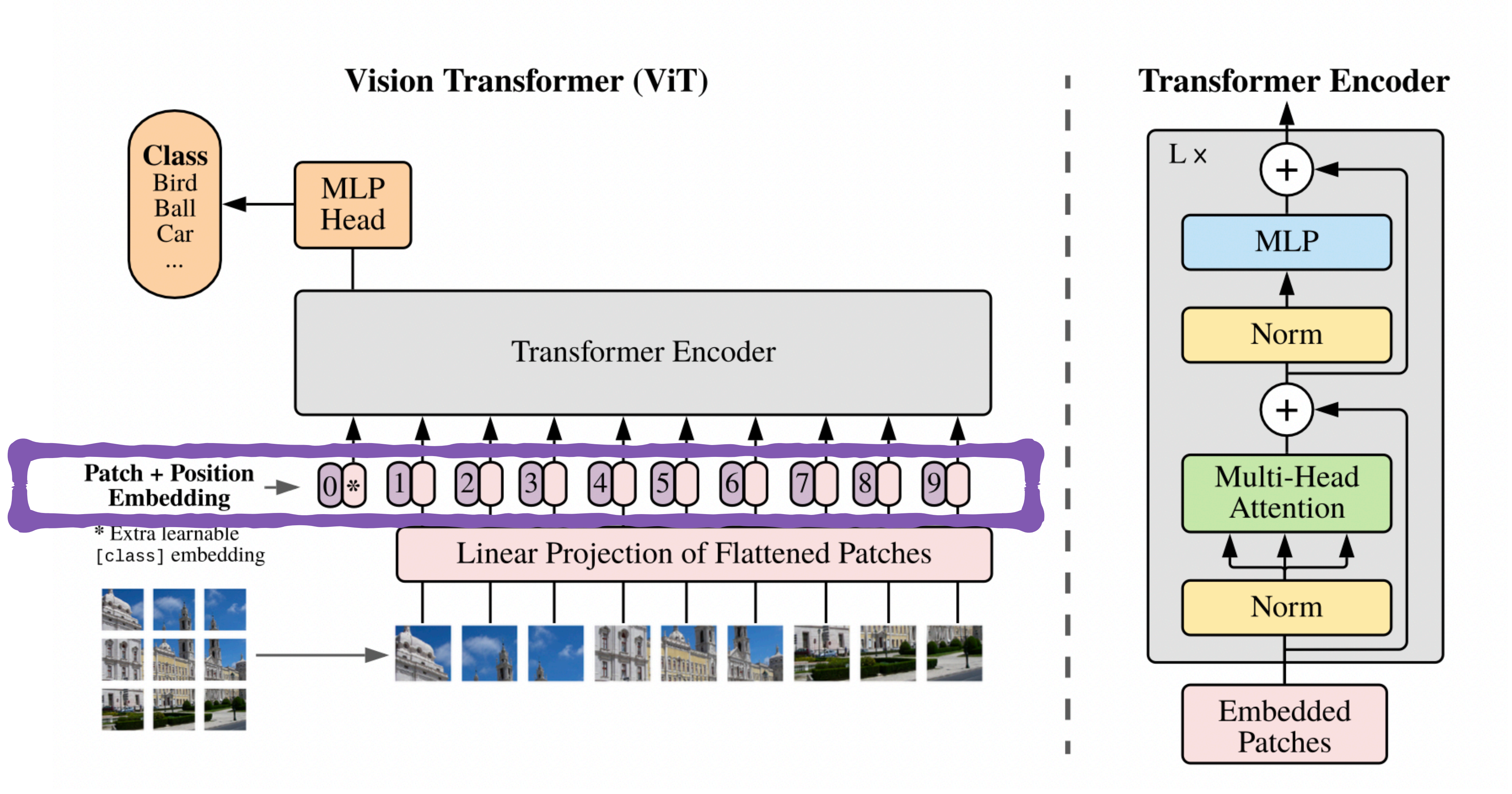
Method

Vision Transformer (ViT)

[CLS] Token

- 임베딩된 패치들의 sequence에 learnable embedding 추가 ($z_0^0 = x_{class}$)

- transformer encoder z_L^0 의 output state : $\mathbf{y} = LN(z_L^0)$



Position embedding

- patch의 위치정보
- 1D Position Embedding 사용

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}, \quad E \in R^{(P^2 \cdot C) \times D}, E_{pos} \in R^{(N+1) \times D}$$

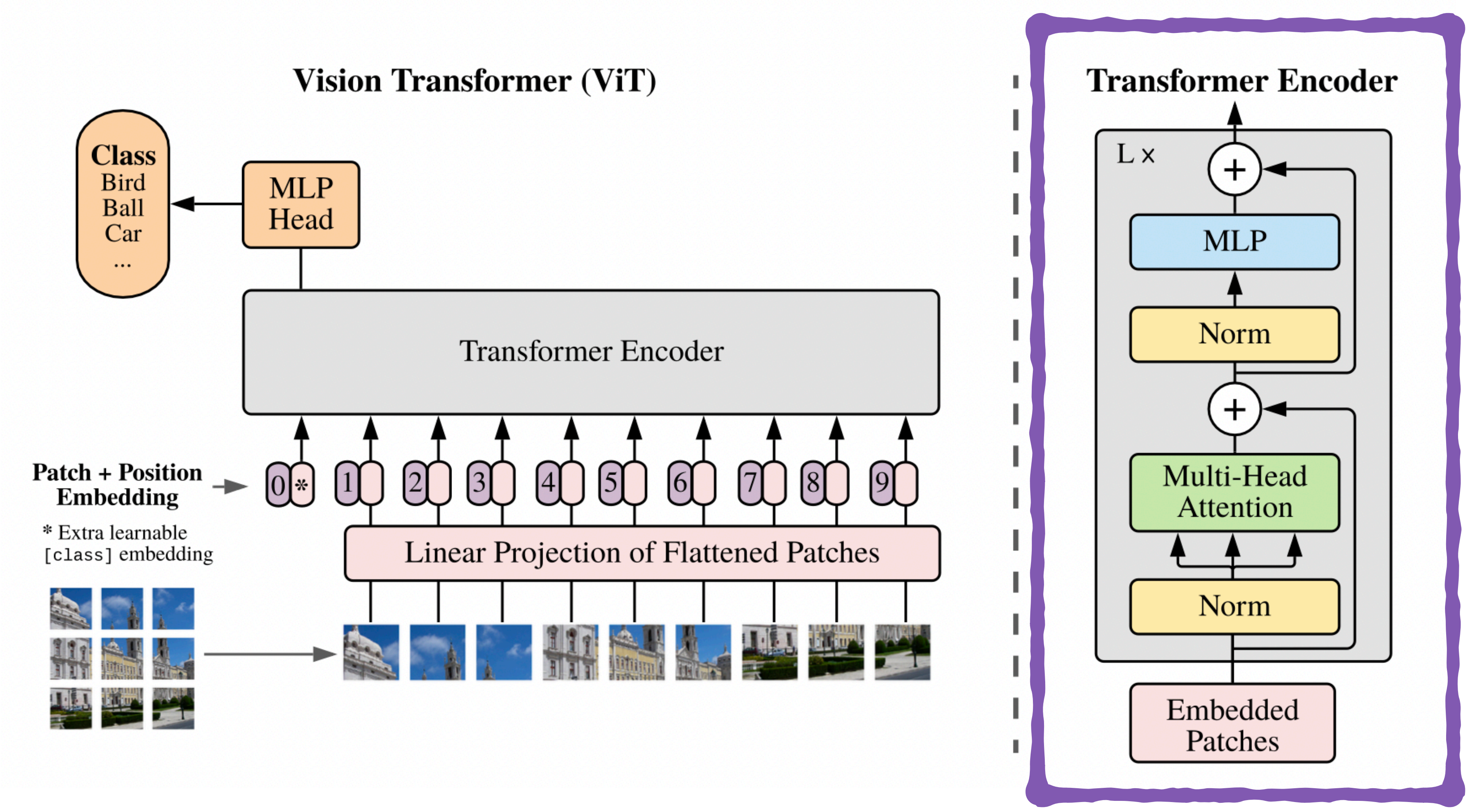
Method

Vision Transformer (ViT)

Transformer Encoder
- layer Norm의 위치가 Transformer 학습에 중요한 역할

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, \qquad l = 1 \ldots L$$

$$z_l = MLP(LN(z'_l)) + z'_l, \qquad l = 1 \ldots L$$



Method

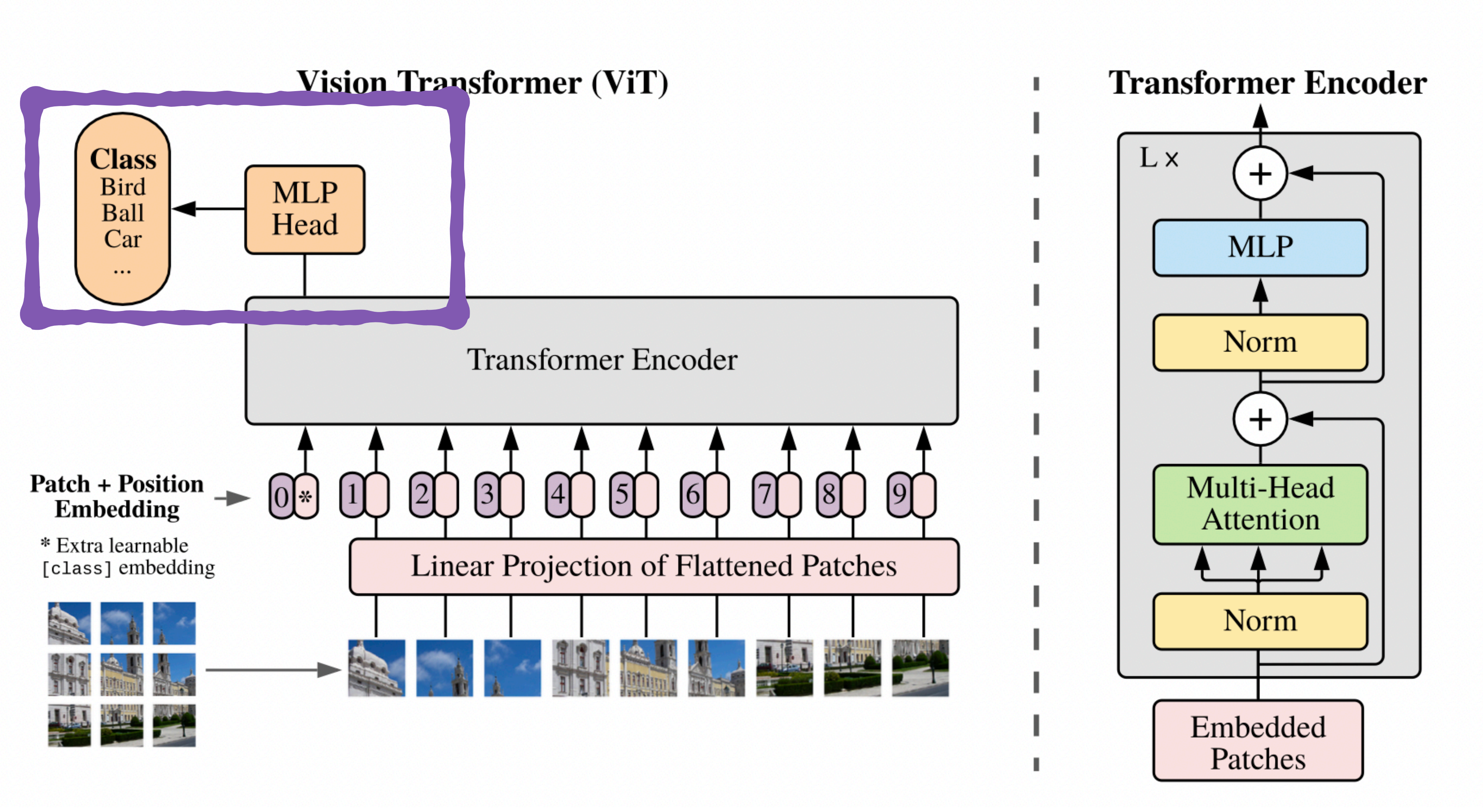
Vision Transformer (ViT)

Output

pre-training : one hidden layer

fine-tuning : one linear layer

$$y = LN(z_L^0)$$



Method

Fine-Tuning & Higher Resolution

- pre-train ViT on large datasets, and fine-tune to (smaller) downstream tasks.
- Pre-training image resolution **<** Fine-tuning image resolution

Experiments

SetUp

[Datasets]

- **Pre-training** - ImageNet-1k(1.3 M), ImageNet-21k(14 M), JFT-18k(303M)
- **Transfer Learning** - ImageNet, ReaL labels, CIFAR 10/100, Oxford-IIIT Pets, Oxford Flowers-102 ..

Experiments

SetUp

[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 : ‘large’ variant with 16x16 input patch size

ResNet(BiT) : Batch Normalization → Group Normalization & Weight Standardization

Experiments

SetUp

[Training]

- Optimizer : Adam($\beta_1 = 0.9, \quad \beta_2 = 0.999$)

- Batch size : 4096

- Weight decay : 0.1

[Fine-Tuning]

- Optimizer : SGD

- Batch Size : 512

- Higher resolution : ViT-L/16 -512 , ViT-H/14: 518

Experiments

COMPARISON TO SOTA

	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

- pre-trained on the JFT-300M
- pre-trained on the smaller public ImageNet-21k

Experiments

COMPARISON TO SOTA

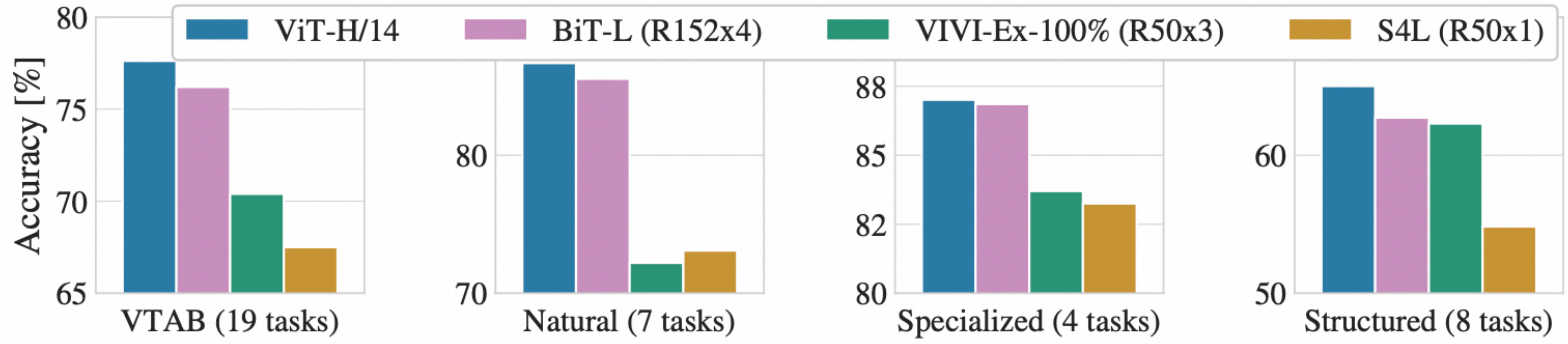
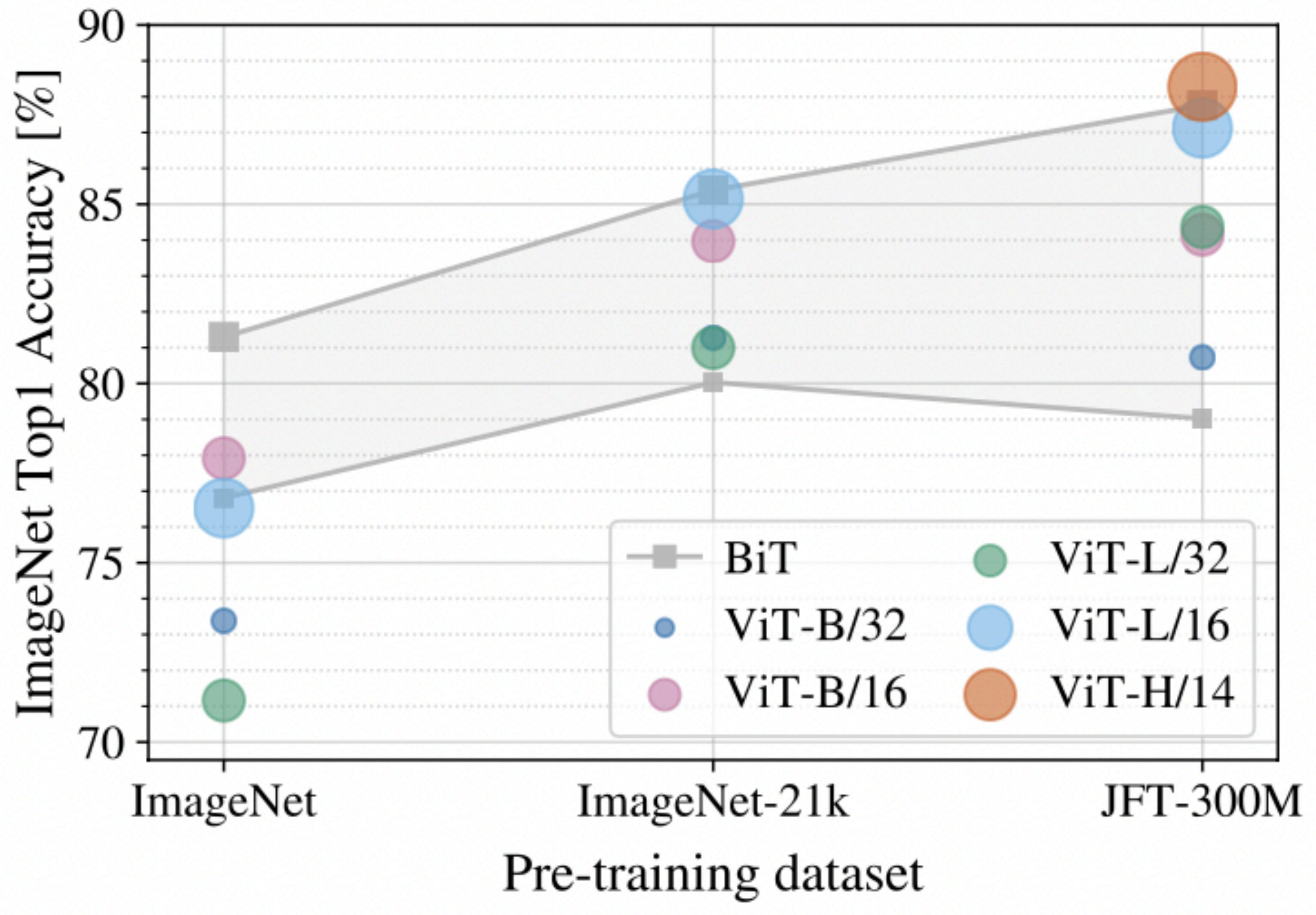


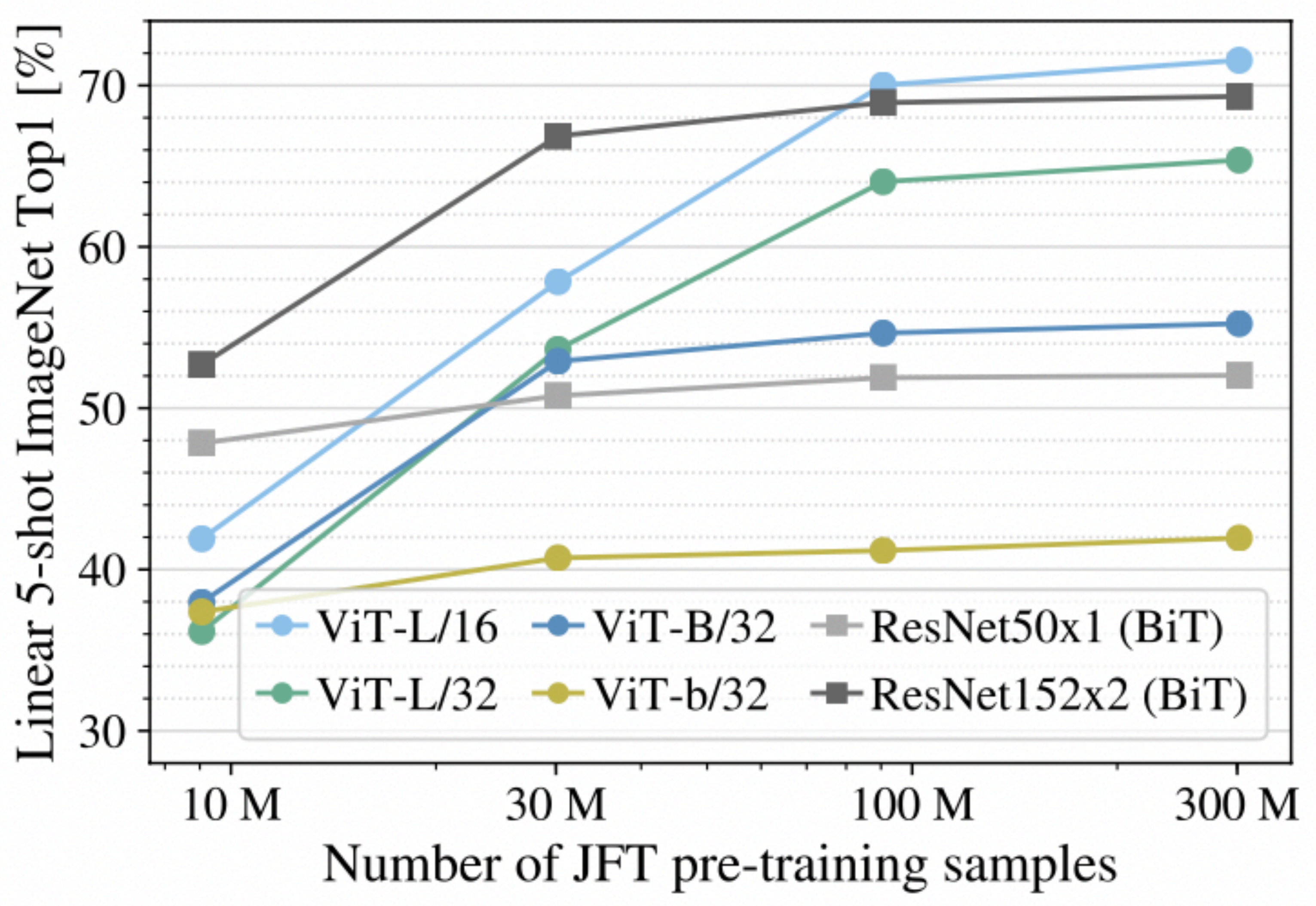
Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

VTAB : Vision Task Adaptation Benchmark

Experiments

Pre-training Data Requirements





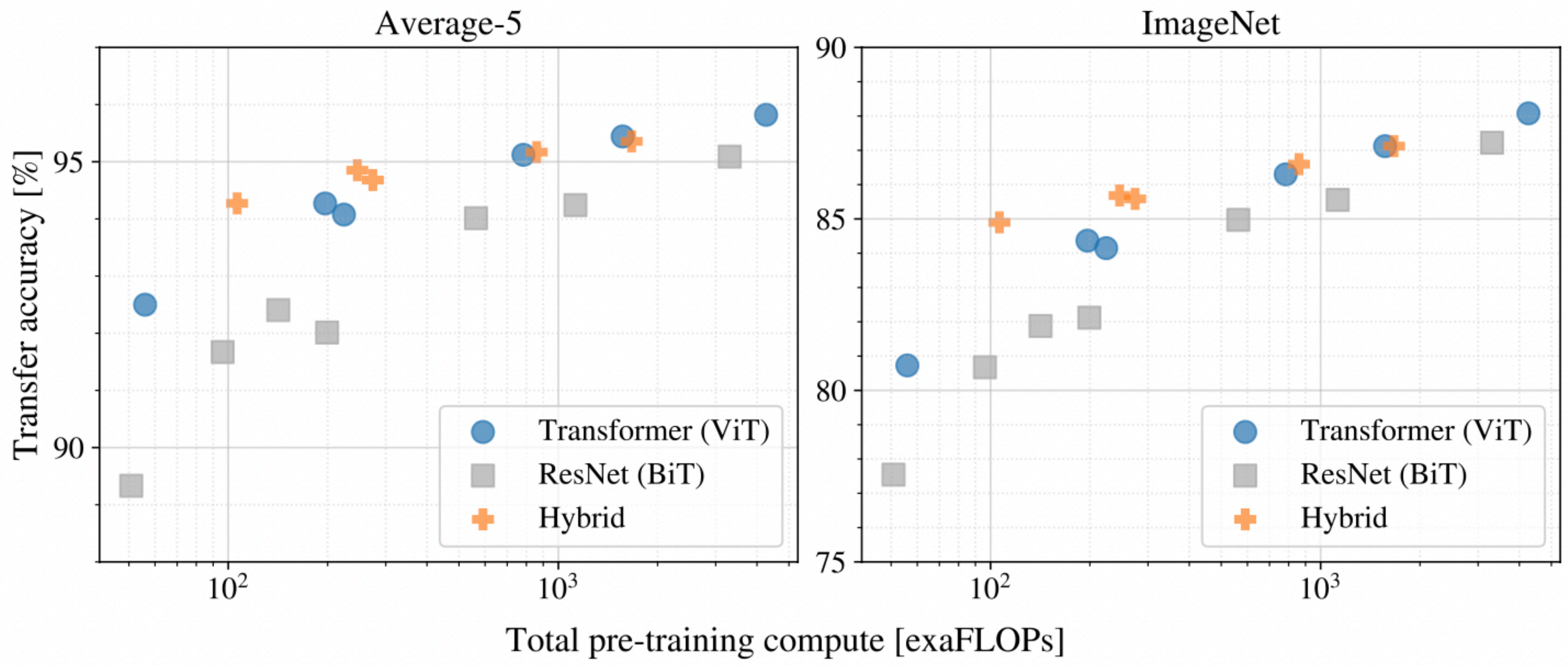
Pre-training : ImageNet, ImageNet-21k, JFT-300M
Fine-Tuning : ImageNet
Regularization :
weight decay, dropout, label smoothing

Small Dataset - Convolutional inductive bias 가 유리

But, 데이터가 많으면 데이터로부터 직접 패턴을 학습하는 것이 유리하다!

Experiments

Scaling Study



- JFT-300M 데이터셋
- ViT는 아직 포화(saturate)되지 않음 → 향후 확장 가능성

Conclusion

- image-specific inductive bias < Large Dataset (JFT-300M)
- Pre-training 비용 저렴

Challenge

- detection 및 segmentation 분야에 적용
- pre-training method
- scaling 통한 추가적인 성능 향상

Thank You