

ICLR 2021

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

2022.08.03

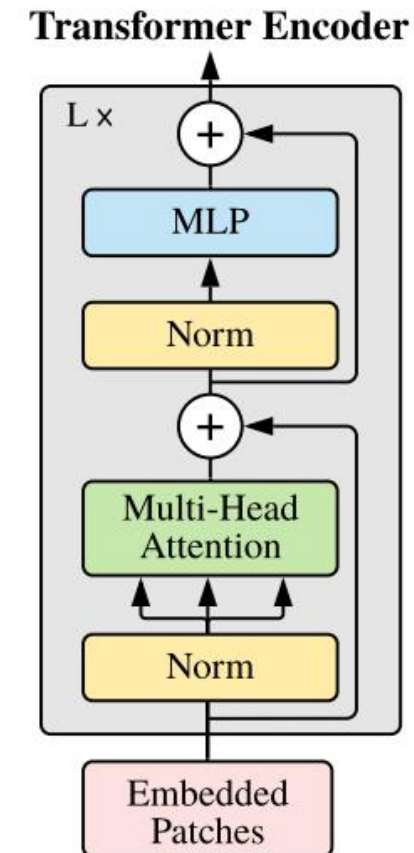
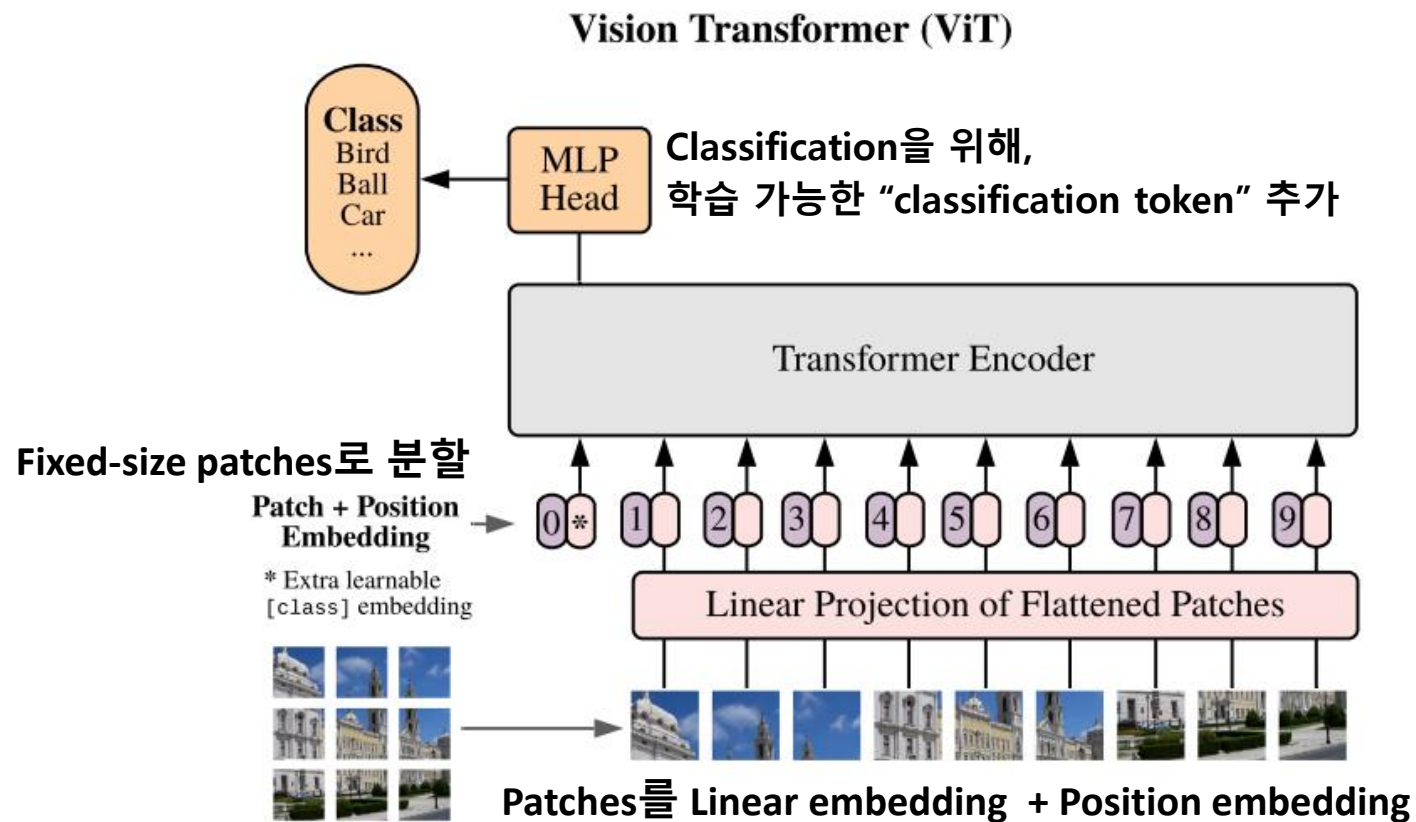
논문 리뷰

배성훈

# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Research Background:**

- 기존의 CV 연구에 NLP의 Transformer를 적용 (CNN 사용 X)
- 그러나 Transformer가 translation equivariance 와 locality 같은 CNN 고유의 일부 inductive biases가 부족해 불충분한 양의 데이터를 학습할 때 쉽지 않은 일반화
- 이를 **large scale training**이 해결 가능. 충분한 양의 사전 학습!



# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method:**

- <Input>

- 일반적인 Transformer는 token embedding에 대한 1차원의 sequence를 입력으로 받음
  - 2차원의 이미지를 다루기 위해 논문은 3차원 이미지를 flatten된 2차원 patch의 sequence로 변환

$$\mathbf{x} \in \mathbb{R}^{\underbrace{H \times W \times C}_{3\text{차원}}} \rightarrow \mathbf{x}_p \in \mathbb{R}^{\underbrace{N \times (P^2 \times C)}_{2\text{차원}}}$$

$(H, W)$  = original image resolution,  $C$  = number of channel

$(P, P)$  = image patch size

$N$  = number of patch =  $HW/P^2$

## Patch + Position Embedding

\* Extra learnable  
[class] embedding

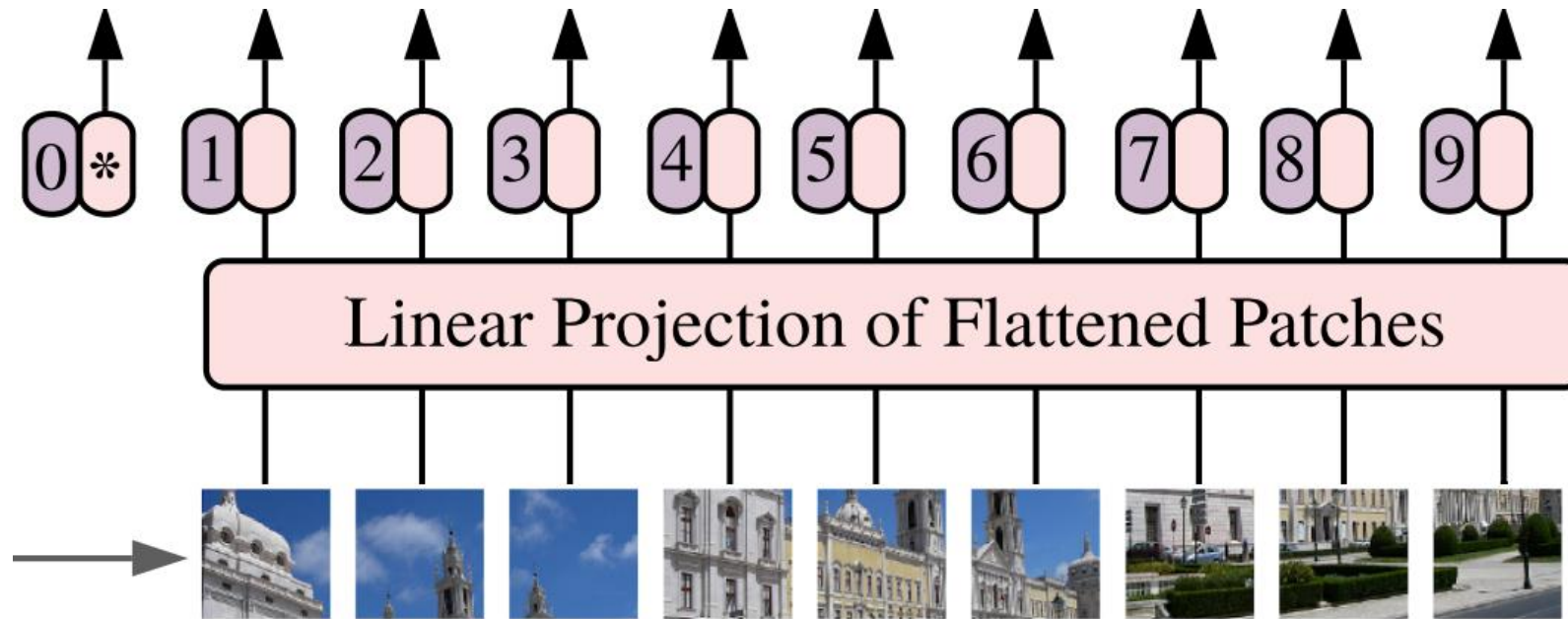


# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method:**

- <Embedding Sequence Patch>

- Transformer는 input으로 **1D sequence of token embeddings**를 가지기 때문에 image patch (2D images)를 Flatten한 후 학습 가능한 linear projection을 사용해 D 차원에 mapping



# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method:**

- <Embedding Sequence Patch>

- BERT의 CLS token과 비슷하게, Transformer encoder ( $z_L^0$ ) 의 output이 image representation  $y$  역할을 하도록 만듦

## Embedding Sequence Patch

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

*Layer norm(LN)*

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

Trainable class token embedding 추가

### Position embeddings

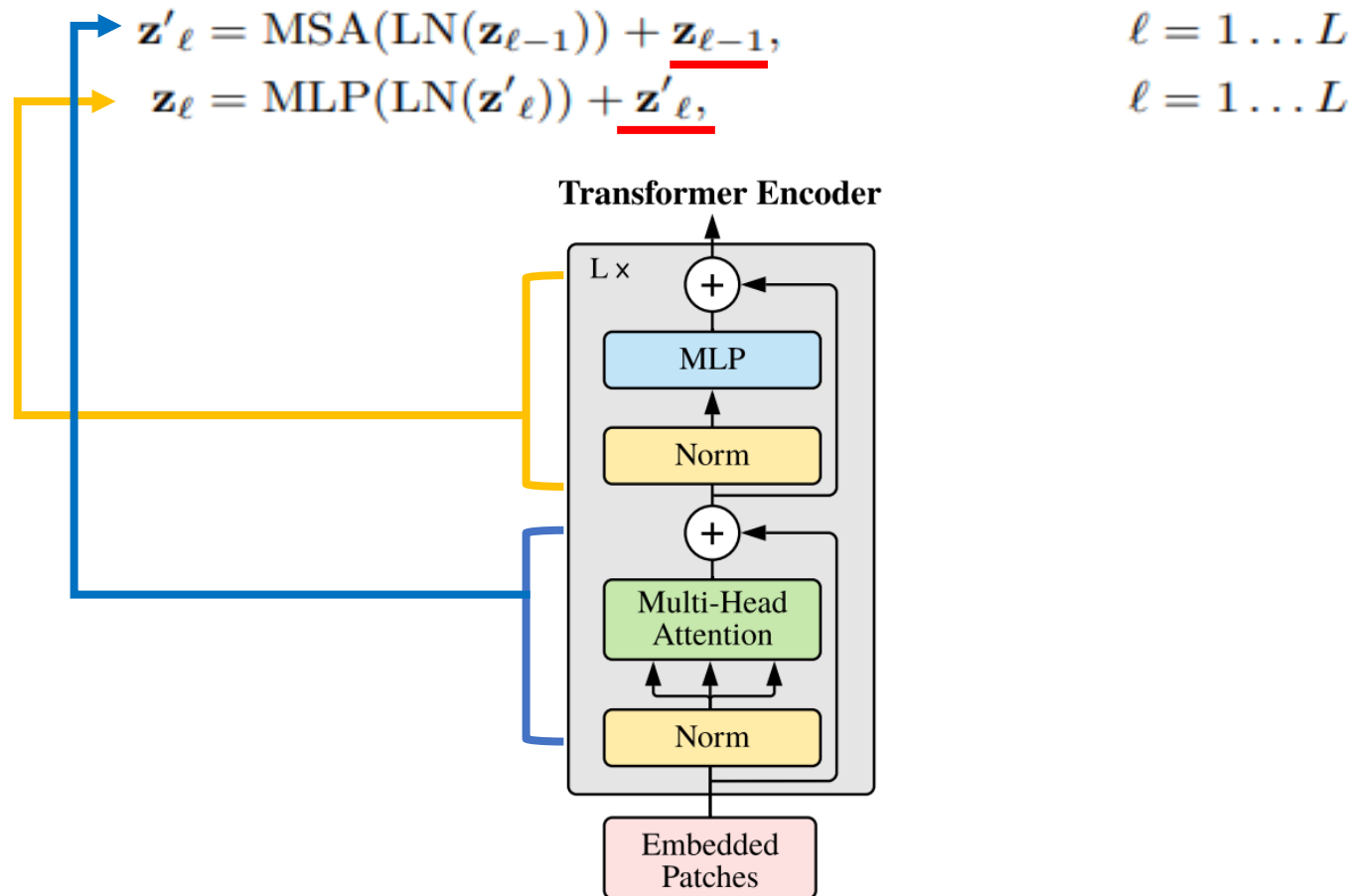
각각의 Patch embedding + Position embedding  
위치 정보를 활용  
학습 가능한 1차원의 embedding을 사용

# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method:**

- <Transformer Encoder>

- **Multihead Self-Attention (MSA) + MLP**
    - Layernorm(LN)은 모든 block 앞에 적용, 모든 block의 뒤에는 Residual connection이 추가
    - 여기서 MLP는 GELU 비선형성을 갖는 2개의 layer를 포함

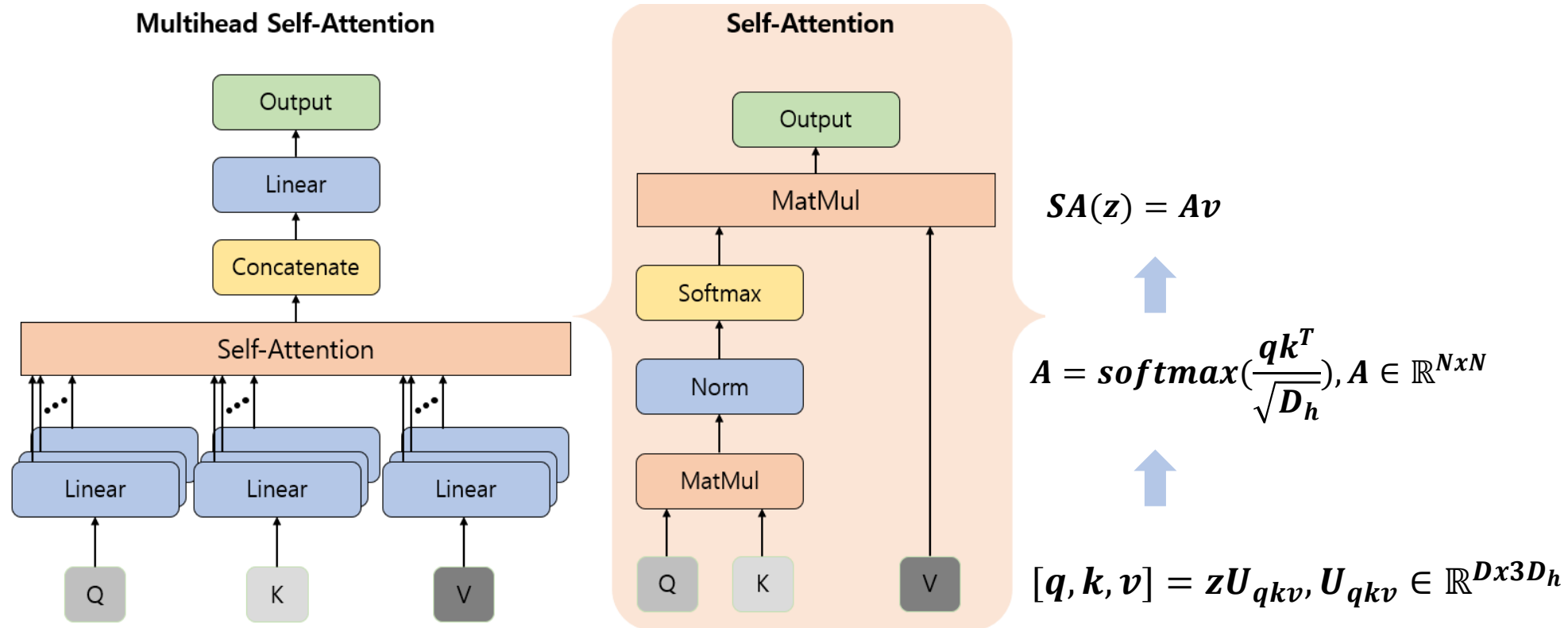


# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method:**

- <Multihead Self-Attention>

Self-Attention 유닛을 Multihead 형태로 확장하기 위해 차원 변경 시 head dimension 반영  
SA 유닛을 k개 concat



$$MSA(z) = [SA_1(z); SA_2(z); \dots; SA_k(z)]U_{msa}, U_{msa} \in \mathbb{R}^{D \times D}$$

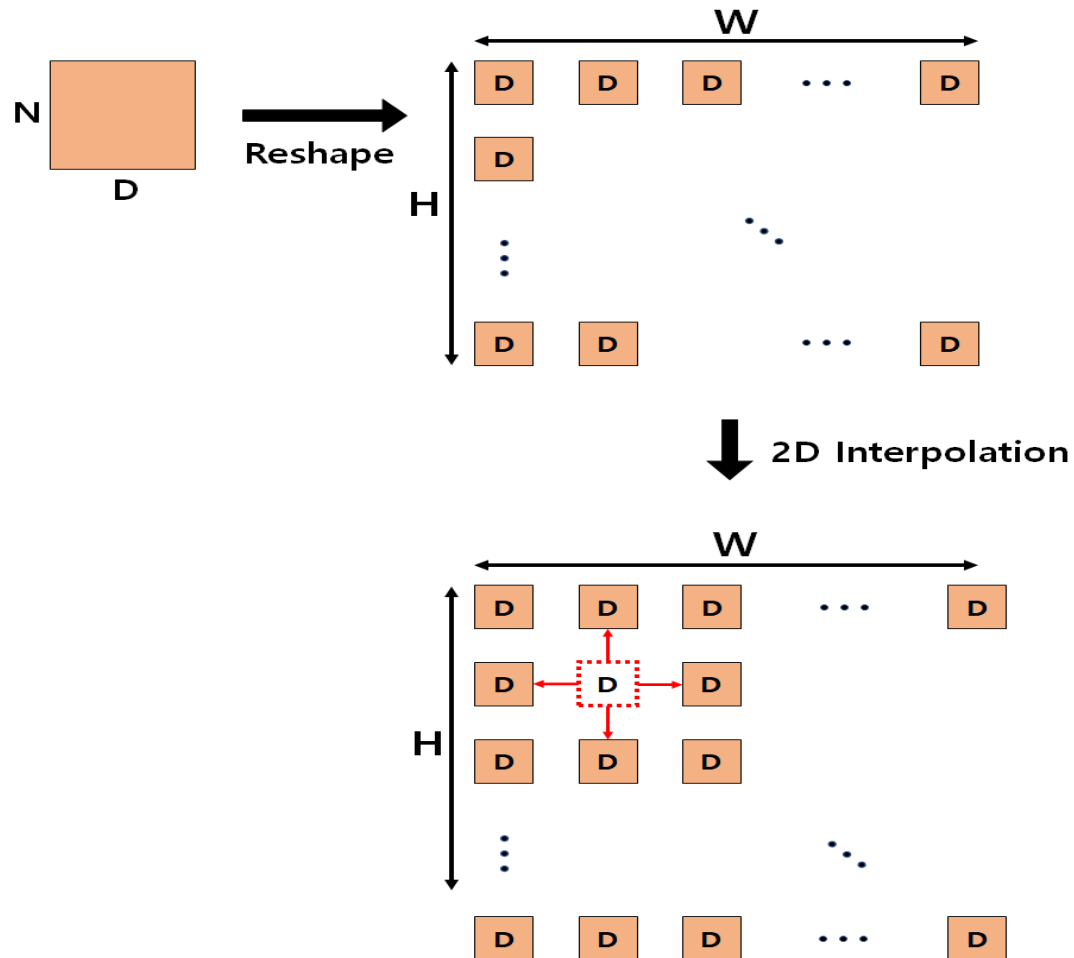
$$\mathbb{R}^{D \times D} = \mathbb{R}^{k \cdot D_h \times D}, D_h = D/k$$

# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method: <Fine tuning and Higher resolution>**

- ViT는 대량의 데이터셋에 대해 사전 학습한 후 더 작은 downstream tasks에 fine-tuning을 하는 방법을 취함
- Fine-Tuning 시 사전 학습된 prediction head를 제거하고, 0으로 초기화된  $D \times K$  FC layer를 부착 ( $K$ = Downstream class의 개수)

## 2D Interpolation when fine tuning



사전학습 시보다 더 높은 해상도의 이미지로 Fine-tuning하는 것이 더 좋은 결과 가져옴

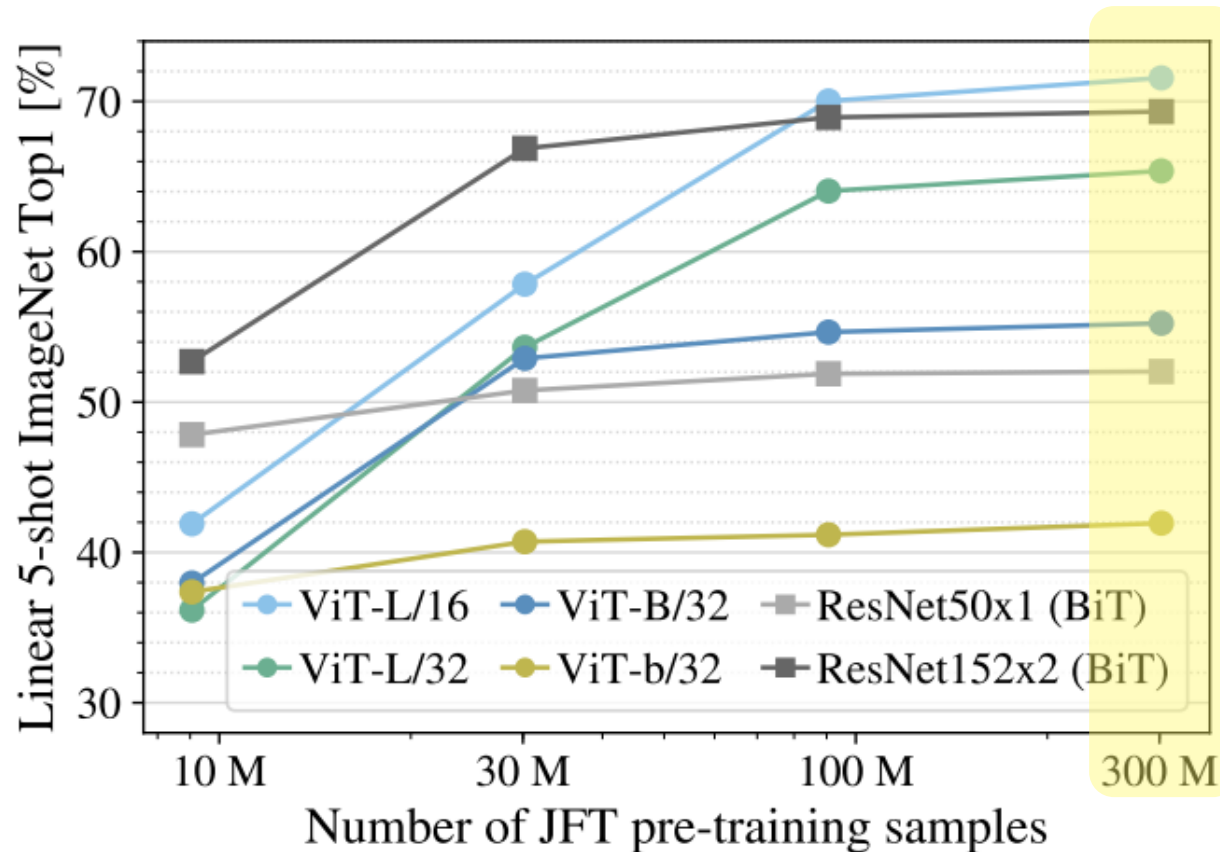


# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Method: <Inductive bias>**

- Self-attention으로 이루어진 ViT는 CNN에 비해 더 작은 image-specific **inductive bias**를 가짐
- ViT는 이를 해결하기 위해 **large datasets**(ex. 14M-300M images)로 **pretrained** 시켜 이를 사용해 specific task with fewer datapoints에 transfer learning을 함

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



Pretraining size ↑ = Performance ↑

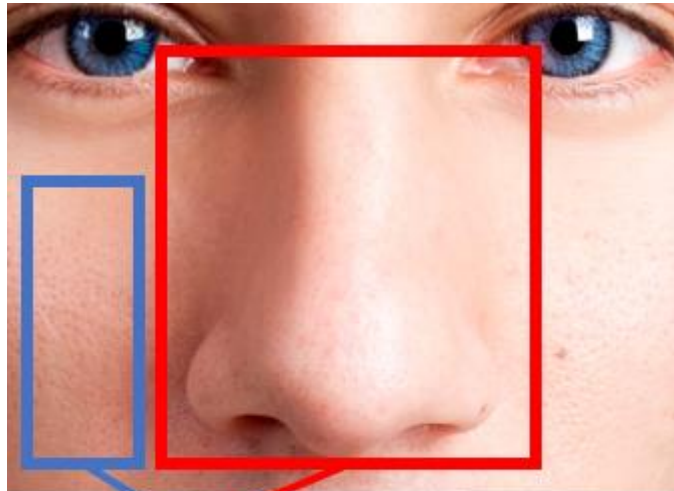
# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- Method: <Inductive bias>

- \*대표적인 Inductive bias

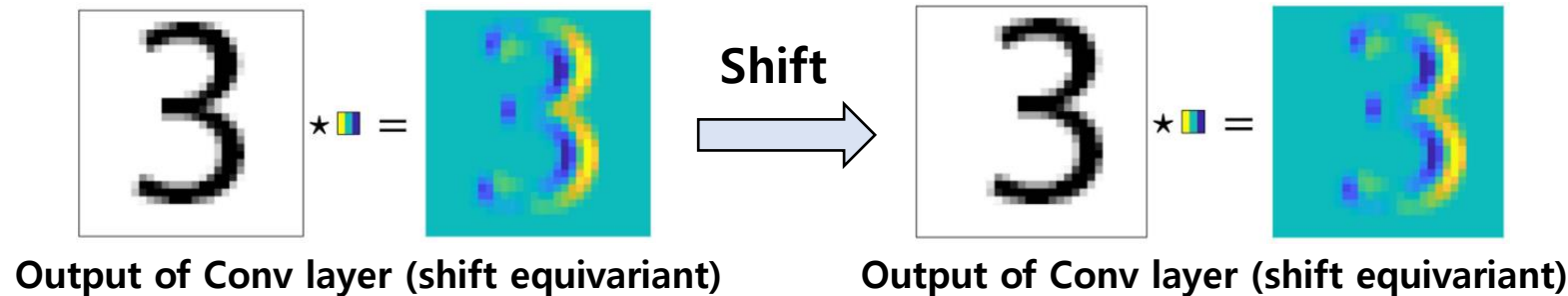
1. **Locality** (Neighborhood 픽셀이 가까울수록 영향도가 커짐)
2. **Translation Invariant** (object가 x, y축으로 이동하거나 회전해도 같은 object 인식)

- 적은 Inductive bias: Optimal parameter를 찾기 위한 space가 커져 데이터가 충분하지 않으면 학습이 잘 안됨



종속성 없음

Locality



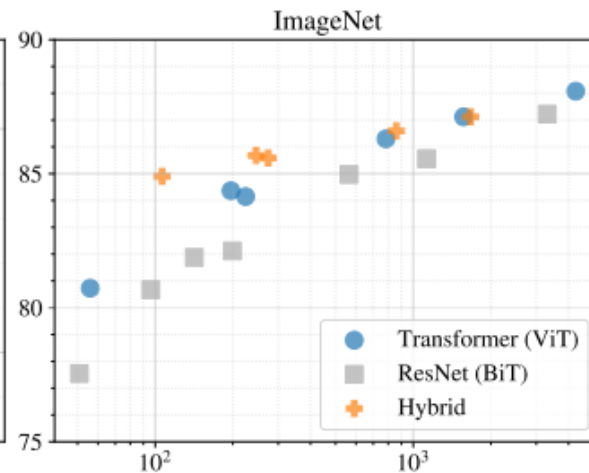
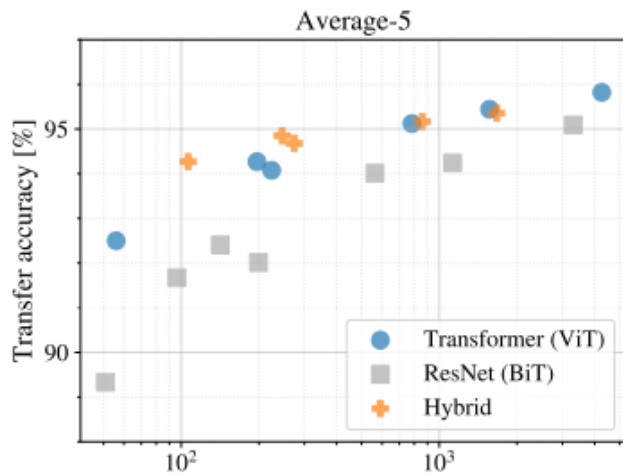
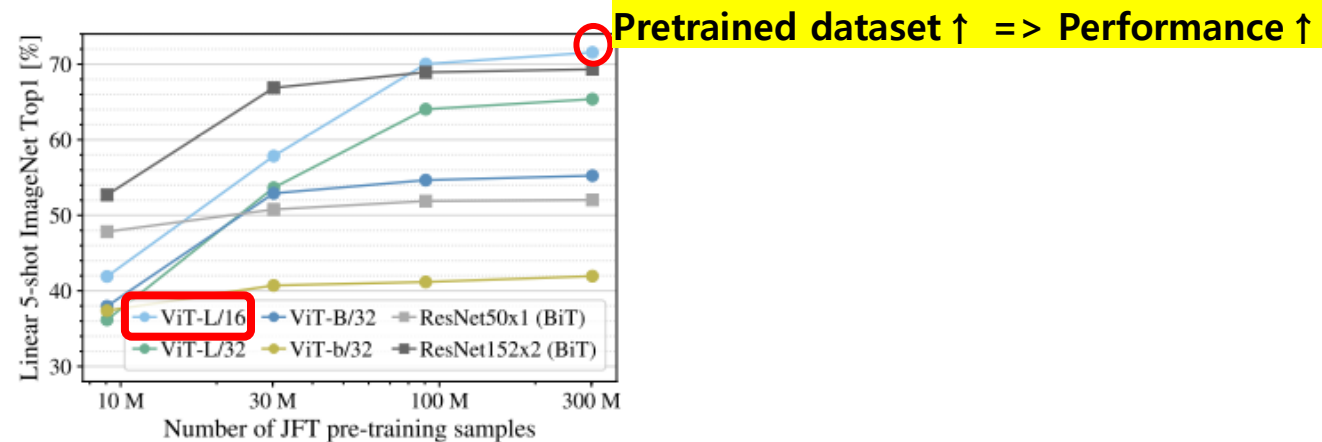
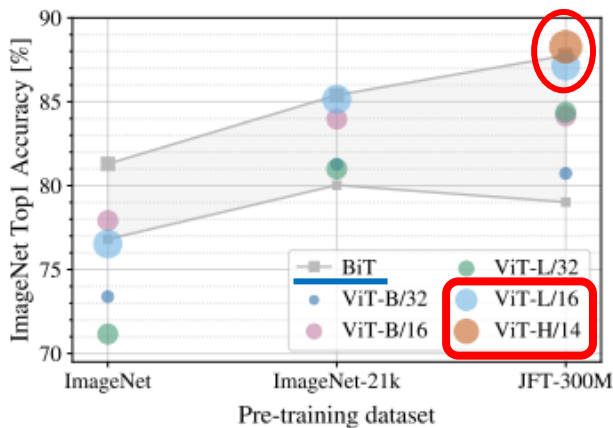
Translation Invariant

# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- Experiment:

- 실험결과, 사전 학습된 데이터가 적을수록 성능이 안 좋아짐
- 즉, ViT는 사전 학습된 데이터가 많아야 좋은 성능을 보임

ViT를 pretraining할 때 사용하는 데이터 셋의 크기에 따른 결과 비교



모델들의 scale을 맞춘 후에 성능 비교 (Scale 지표: FLOPS)

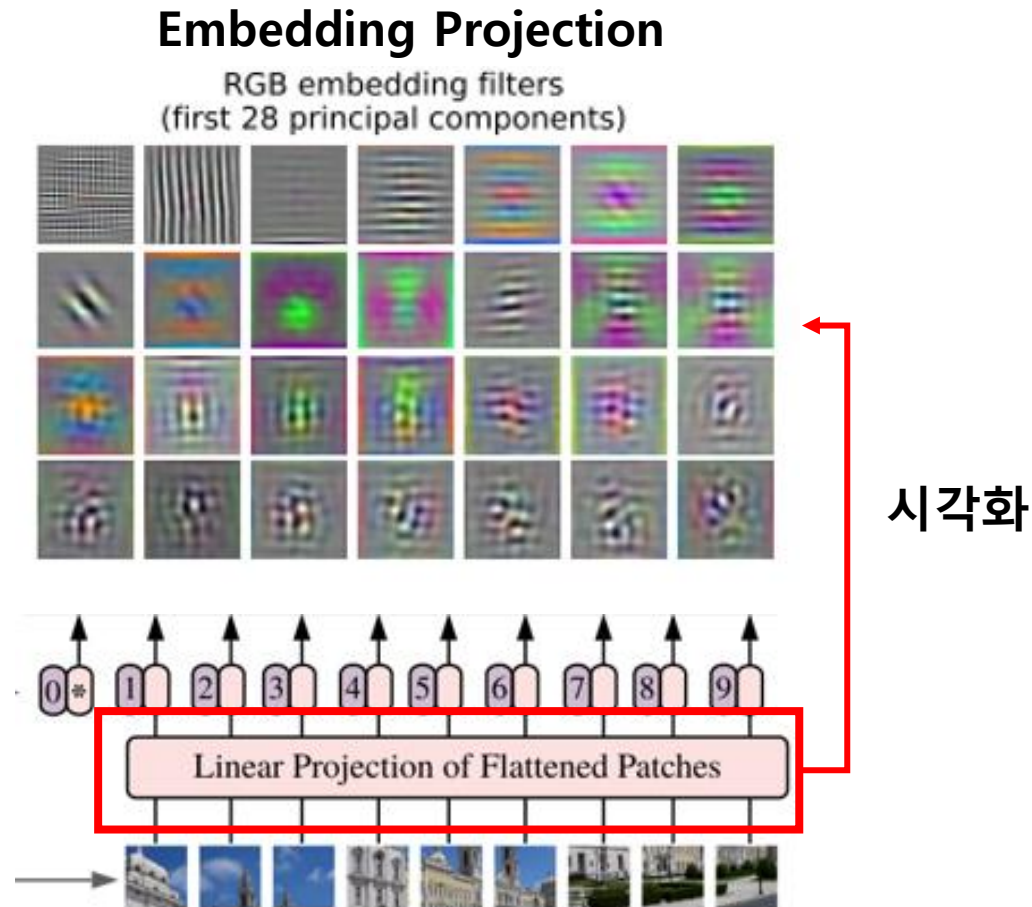
ViT vs ResNet(BiT)

Better Performance/compute trade off

# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Experiment:**

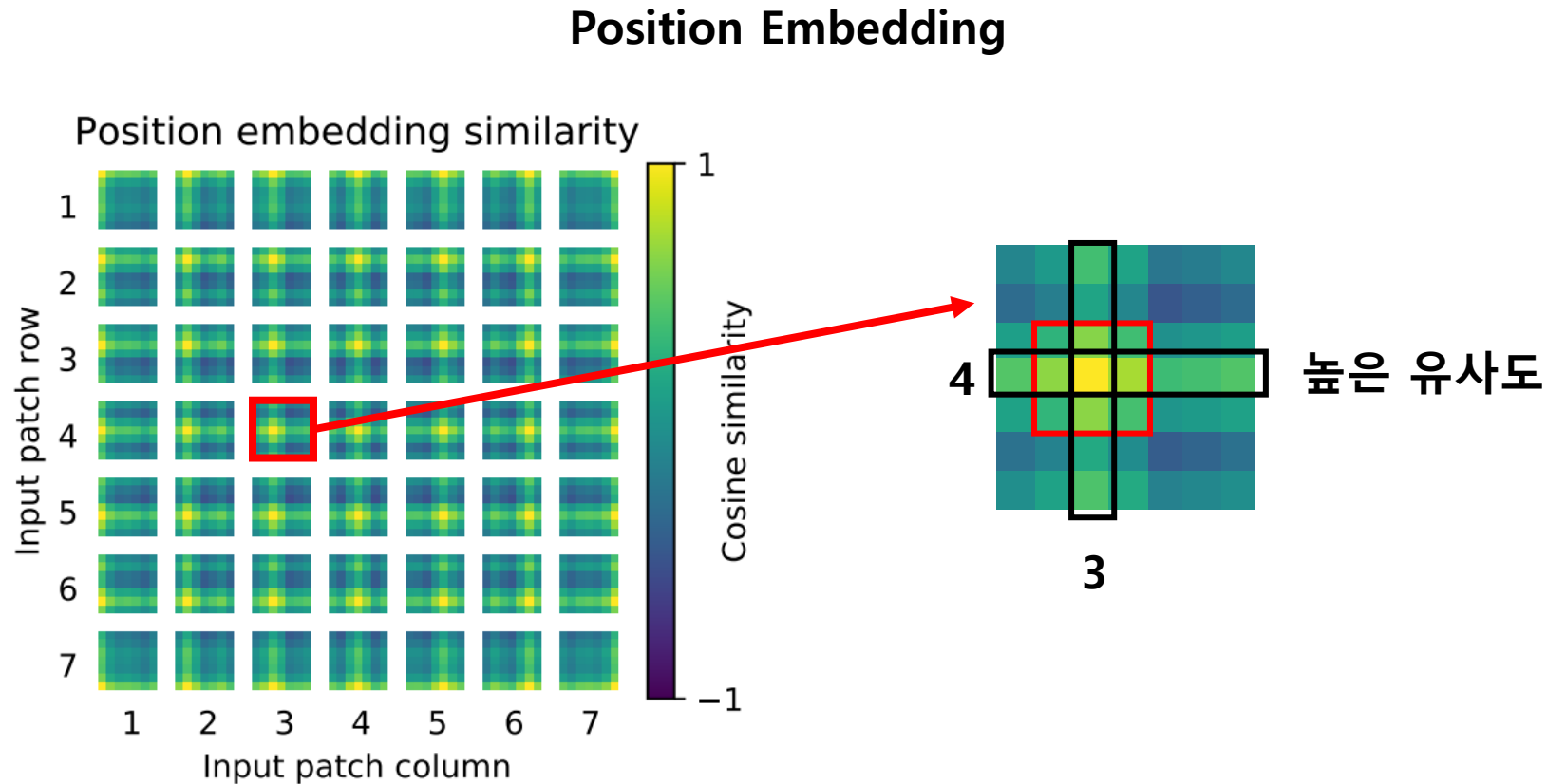
- 첫 번째 Linear Projection 부분에서 주요 요소 28개를 선정해 시각화
- Embedding filter를 시각화 했을 때 CNN filter 와 비슷한 기능을 보임 (많은 데이터를 사전학습한 경우)
- CNN과 같이 이미지 인식에 필수인 Edge, Color 등의 low-level feature들을 잘 포착



# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- Experiment:

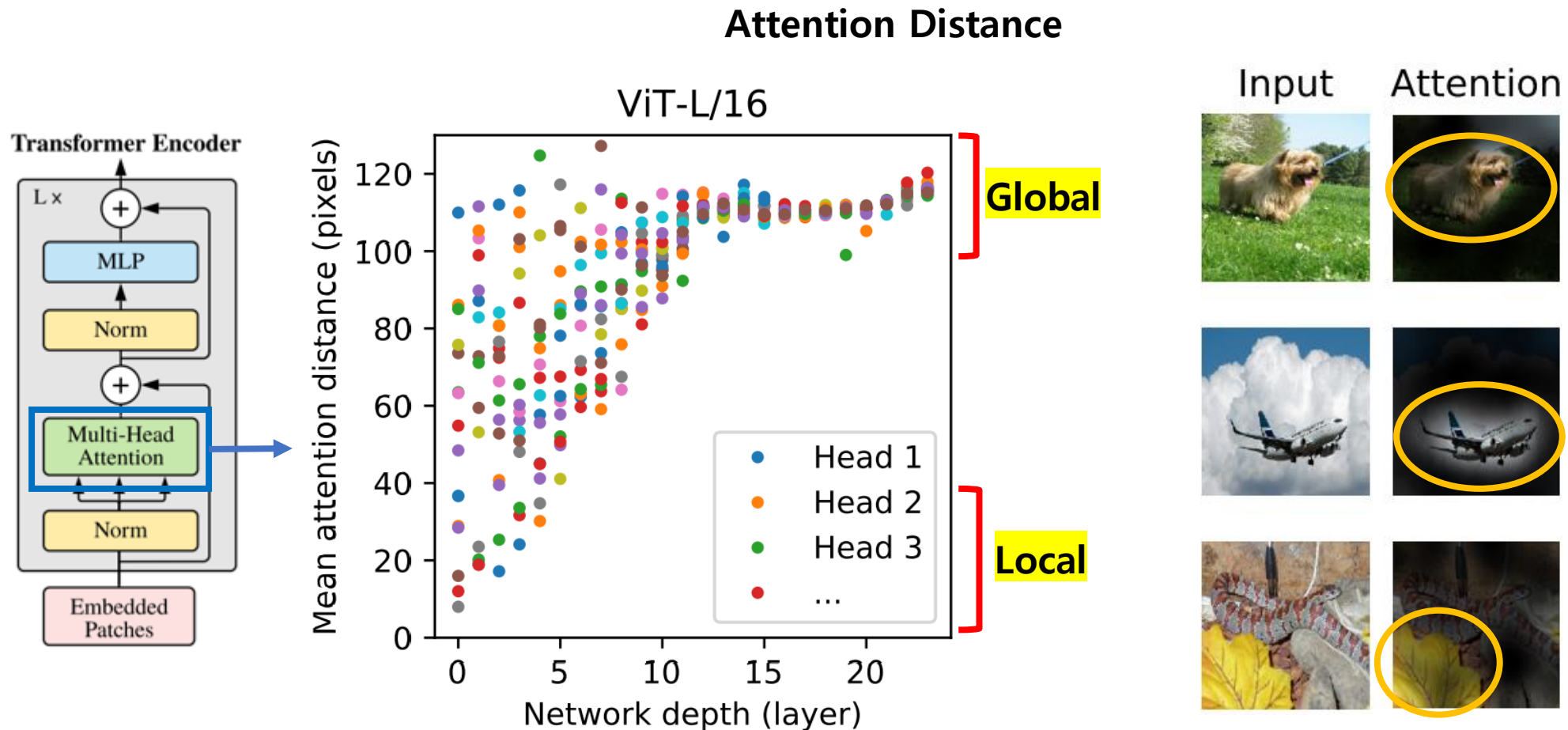
- 가까운 거리, 같은 열 또는 행에 위치한 Patch는 비슷한 position embedding과 높은 유사도를 보임



# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

- **Experiment:**

- Attention Distance는 CNN의 Receptive Field와 유사한 개념
- Self-Attention은 네트워크가 가장 첫 번째 layer에서도 이미지의 Global features 파악
- Attention Distance를 시각화한 결과 첫 번째 layer에서부터 **Local, Global features** 잘 포착





# AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ICLR 2021)

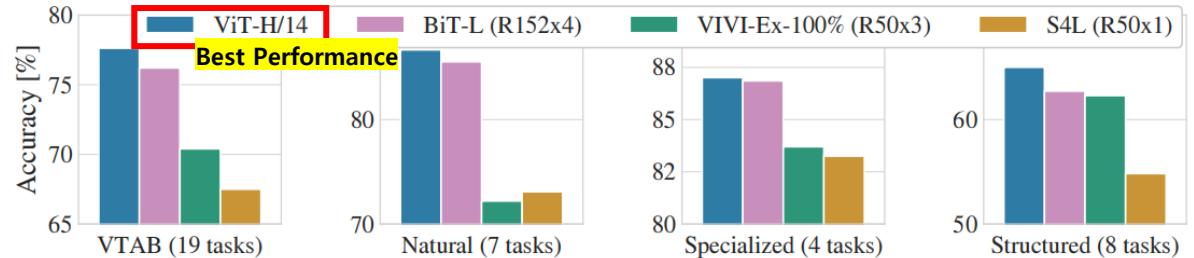
- Experiment:

- 전체적으로 다양한 benchmark dataset 에서 SOTA 달성.

Each Model 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

VTAB performance in Natural, Specialized, and Structured task groups



벤치마크된 lager 데이터셋으로 사전학습한 ViT와 SOTA 모델들의 데이터셋에 따른 성능 비교

	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

한줄평: 개인적으로, 결국 많은 양의 데이터를 사용하지 않는 이상 좋은 성과를 볼 수 없고, 모델을 활용하는데 있어 Google research의 사전학습 데이터를 사용하는 것이 불가피하기 때문에 (Inductive bias 때문) 활용하기 힘든 단점을 가진 것 같다.