

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

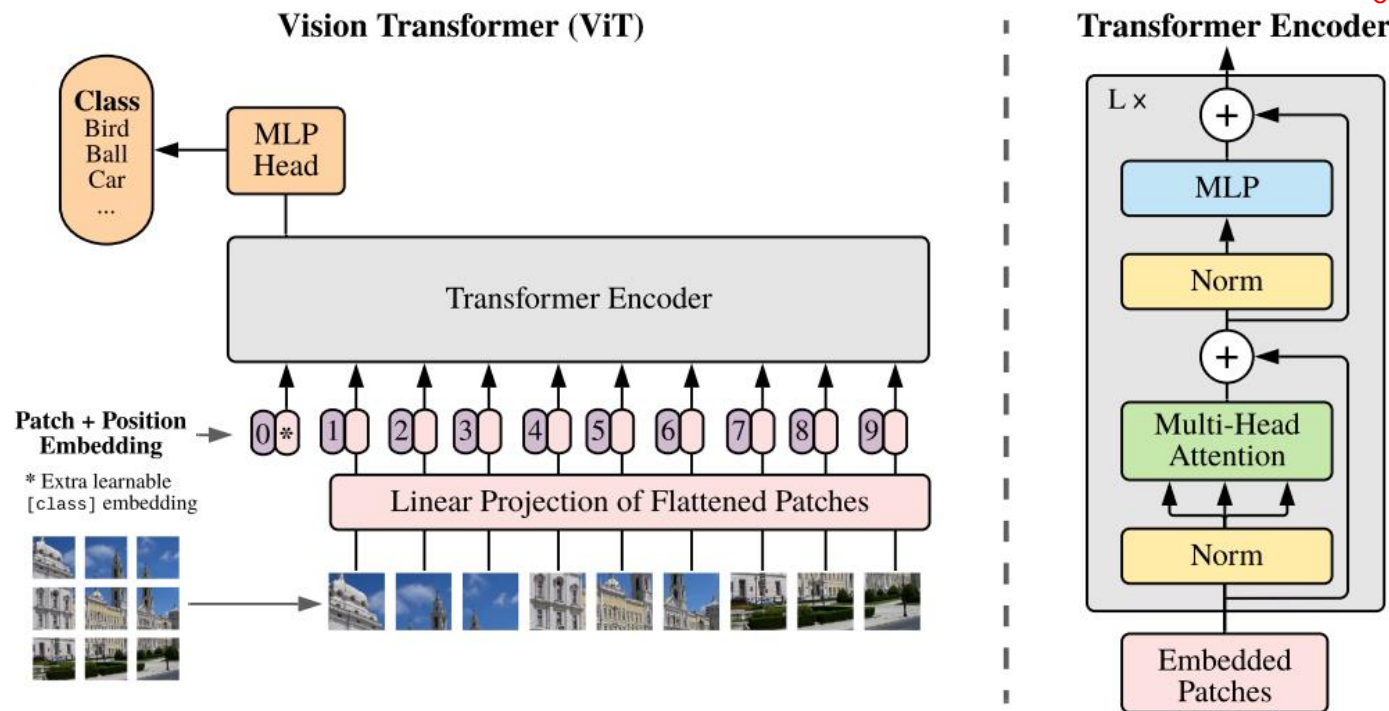
- 연구동기:

기존의 CV 연구에 NLP의 Transformer를 적용 (CNN 사용 X)

문제: Transformer가 translation equivariance 와 locality 같은 CNN 고유의 일부 inductive biases가 부족해 불충분한 양의 데이터를 학습할 때 일반화가 쉽지 않다.

해결: 이를 **large scale training**이 해결 가능. 충분한 양의 사전 학습!

최대한 Standard transformer를 그대로 이미지에 적용하고자 함.



이미지를 fixed-size patches로 분할, 각 patches를 선형으로 embedding + position embedding
생성된 sequence of vectors -> Transformer encoder 공급

Classification을 위해, 학습 가능한 "classification token"을 sequence에 추가

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

- **Method:**

<Input>

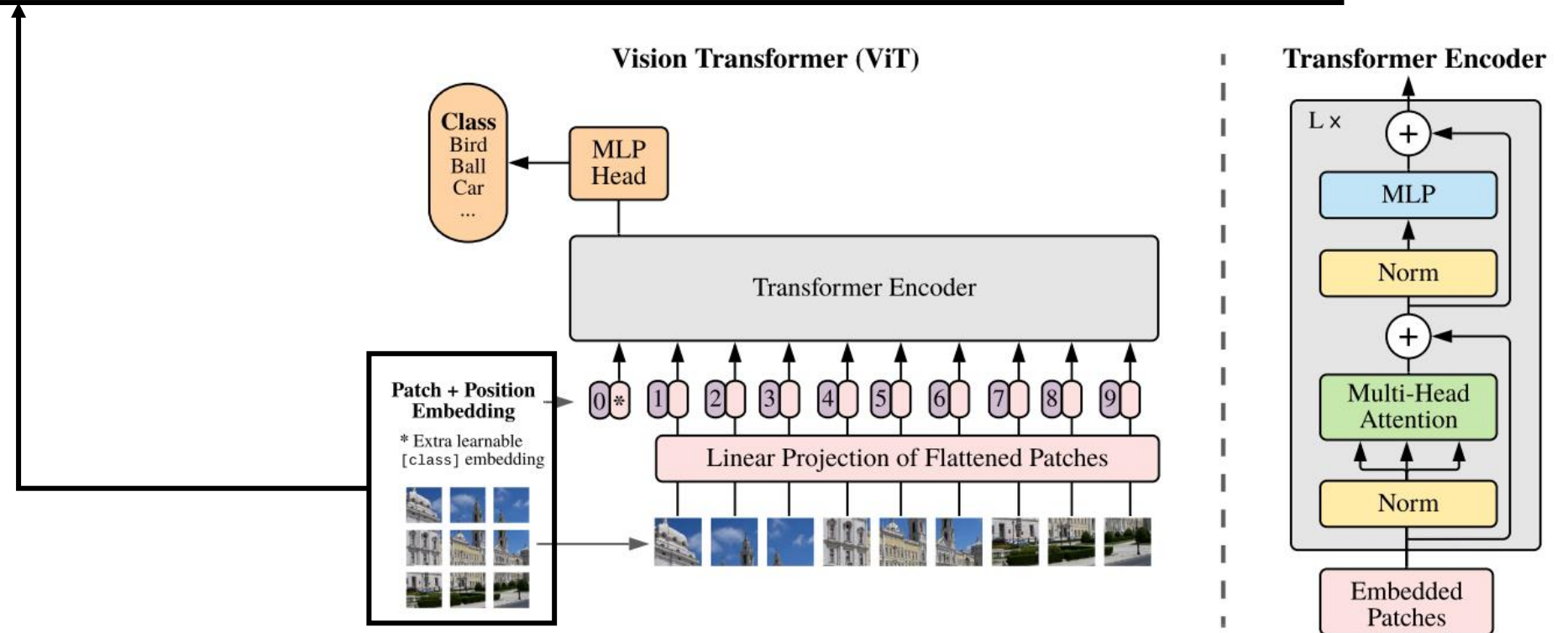
일반적인 Transformer는 token embedding에 대한 1차원의 sequence를 입력으로 받는다.
2차원의 이미지를 다루기 위해 논문은 이미지를 flatten된 2차원 patch의 sequence로 변환

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

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

(P, P) = image patch size

N = number of patch = HW/P^2



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

- Method:

<Standard Transformer>

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

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \cdots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}},$$

Learnable embedding 추가 **Embedding sequence patch**

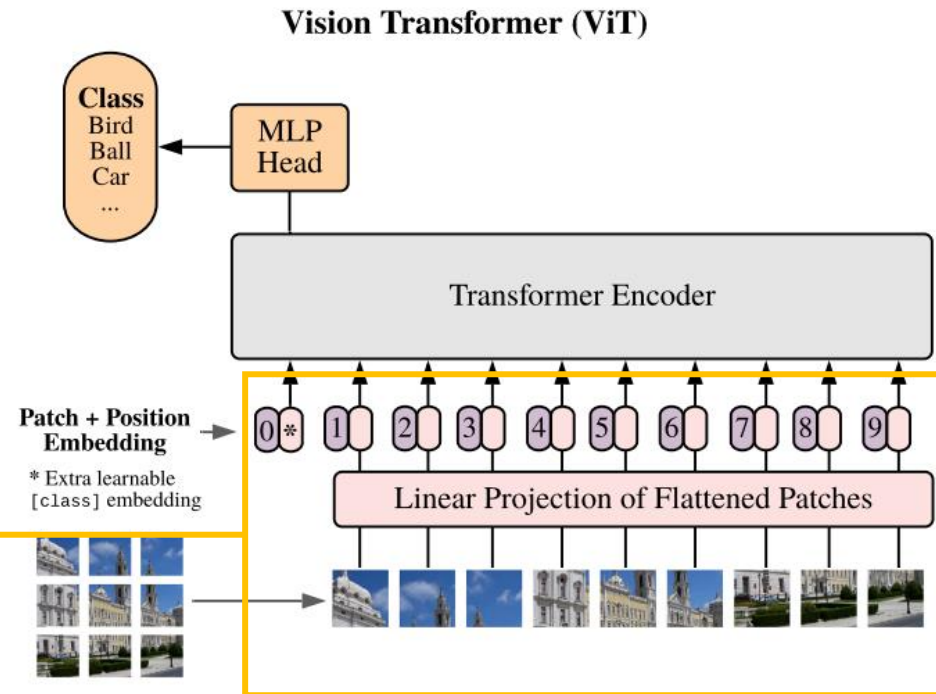
$$\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$$

Position embeddings

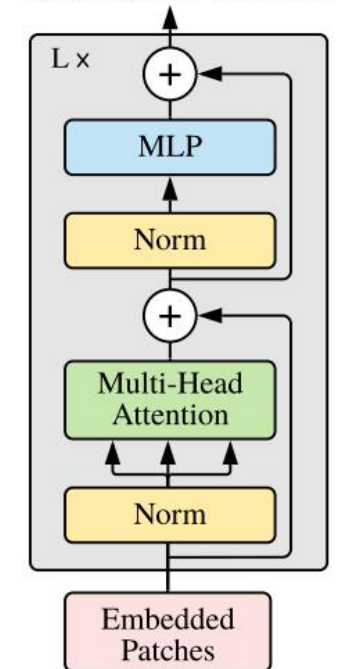
각각의 patch embedding + position embedding => 위치 정보를 활용 학습 가능한 1차원의 embedding을 사용

BERT의 CLS token과 비슷하게, Transformer encoder (\mathbf{z}_L^0)의 output이 image representation \mathbf{y} 역할을 하도록 만든다.

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0), \text{Layer norm}(\text{LN})$$



Transformer Encoder



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

- **Method:**


<Transformer Encoder>

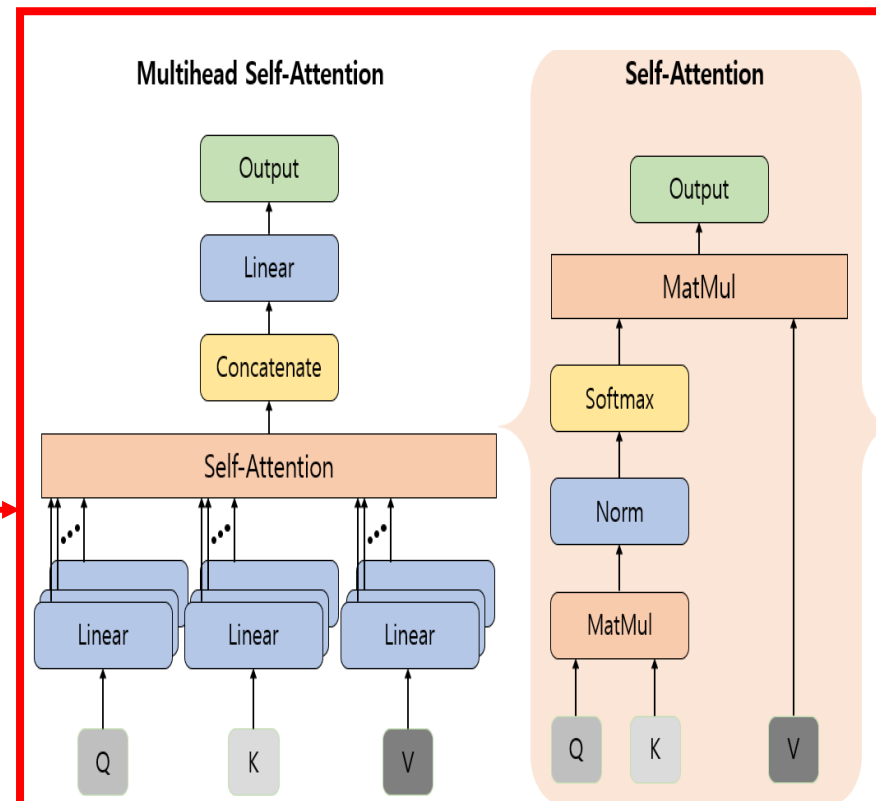
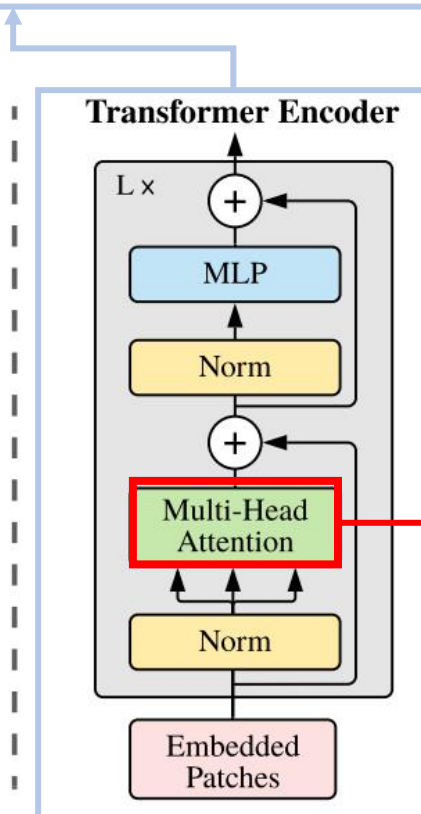
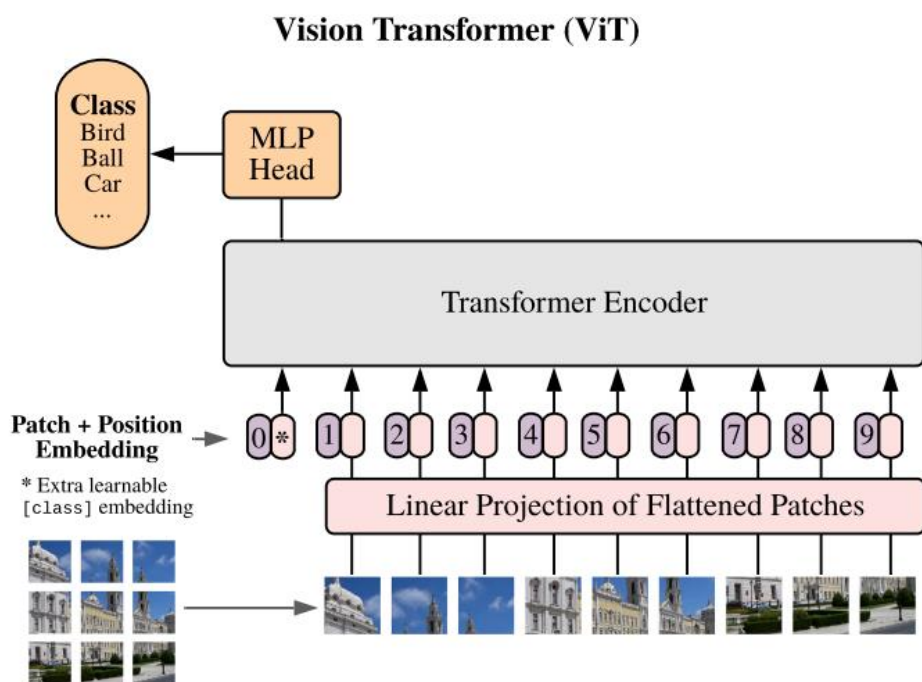
Multihead Self-Attention (MSA) + MLP

Layernorm(LN)은 모든 block 앞에 적용되고, 모든 block의 뒤에는 **residual connection**이 추가된다.

여기서 MLP는 GELU 비선형성을 갖는 2개의 layer를 포함한다

$$\begin{aligned} \mathbf{z}'_{\ell} &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, & \ell &= 1 \dots L \\ \mathbf{z}_{\ell} &= \text{MLP}(\text{LN}(\mathbf{z}'_{\ell})) + \mathbf{z}'_{\ell}, & \ell &= 1 \dots L \end{aligned}$$

 **Residual connection**



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 부착한다.

(K= Number of downstream classes)

이때 fine-tuning단계에서는 더 높은 해상도에서 학습하는 것이 정확도 향상에 좋다는 것으로 알려져 있다.

더 높은 해상도의 이미지를 처리해야 할 경우, image patch size를 동일하게 유지함으로써 더 긴 patch sequence를 사용한다.

ViT는 더 높은 하드웨어의 메모리가 허용하는 한, 임의의 길이의 sequence를 처리할 수 있다.

하지만, 이 경우 사전 학습된 position embedding은 큰 효과를 가지지 못한다.

사전 학습된 position embedding에 원본 이미지에서의 position에 따라 2D Interpolation을 수행한다.

해상도 조절과, patch 추출 방법은 ViT에서 이미지의 2차원 구조에 대한 inductive bias를 수동적으로 다루는 유일한 포인트이다.

<Inductive bias>

Self-attention으로 이루어진 ViT는 CNN에 비해 더 작은 image-specific inductive bias를 가진다.

대표적인 Inductive bias

1. locality (Neighborhood 픽셀이 가까울수록 영향도가 커진다.)
2. translation invariant (object가 x, y축으로 이동하거나 회전해도 같은 object 인식)

적은 Inductive bias = **장**: Constraint 없이 이미지 전체에서 정보를 얻을 수 있다.

단: Optimal parameter를 찾기 위한 space가 커져 데이터가 충분하지 않으면 학습이 잘 안됨

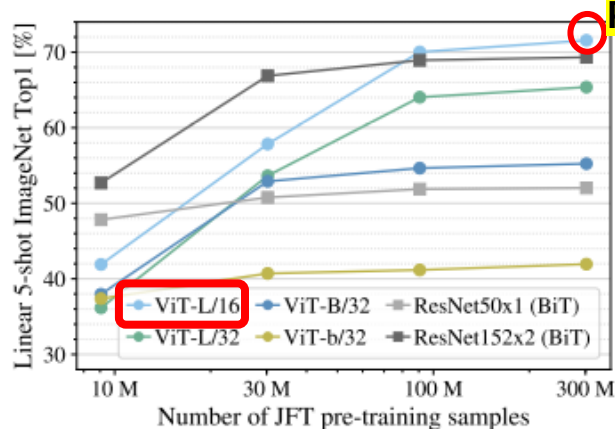
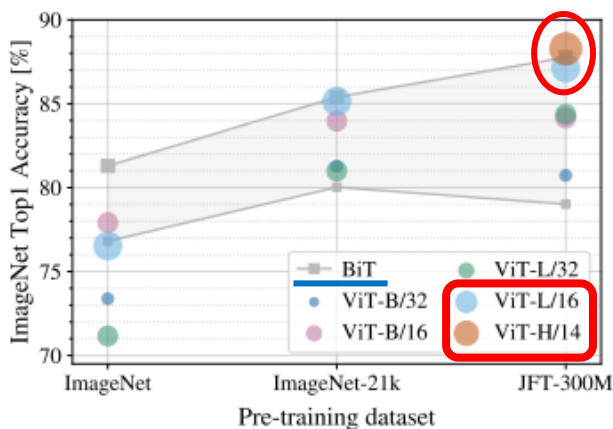
ViT는 이러한 단점을 **large datasets**(ex. 14M-300M images)로 **pretrained** 시켜 이를 사용해 specific task with fewer datapoints에 transfer learning을 한다.

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

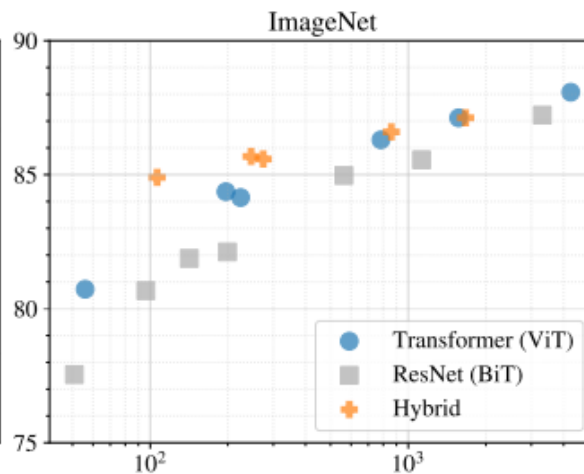
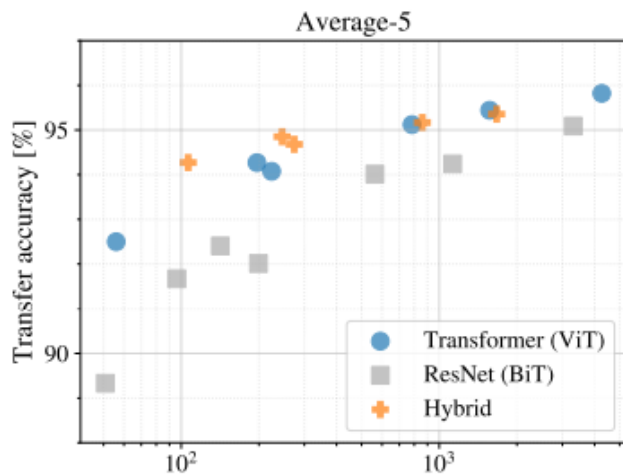
- **Experiment:**

실험결과, 사전 학습된 데이터가 적을수록 성능이 안 좋아진다.
즉, ViT는 사전 학습된 데이터가 많아야 좋은 성능을 보인다.

ViT를 pretraining할 때 사용하는 데이터 셋의 크기에 따른 결과 비교



Pretrained dataset \uparrow \Rightarrow Performance \uparrow



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

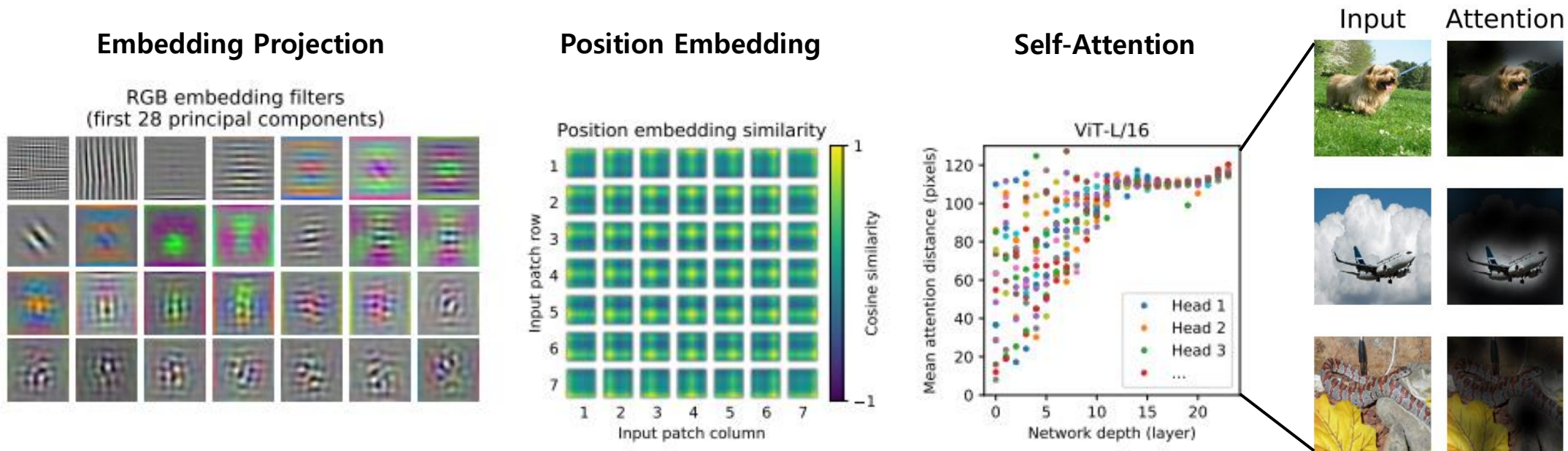
ViT vs ResNet(BiT)

Better Performance/compute trade off

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

- **Experiment:**

Embedding filter를 시각화 했을 때 CNN filter 와 비슷한 기능을 보인다. (많은 데이터 사전학습한 경우)



각각의 Patch에 대해
low dimension representation을
만드는 기본 함수들을 나타냄
= CNN filter와 비슷한 기능

가까운 거리, 같은 열 또는 행에 위치한
Patch는 비슷한 position embedding을
가짐

시각화 정보:

Attention weigh에 기반하여 이미지 공간 상에서
정보가 취합되는 평균 거리 (= CNN의 receptive field)

일부 attention head:

Global하게 정보를 통합하는 능력을 모델이 활용
일관적으로 작은 거리의 patch에 집중
= **CNN에서 발생하는 작용과 비슷하다.**

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

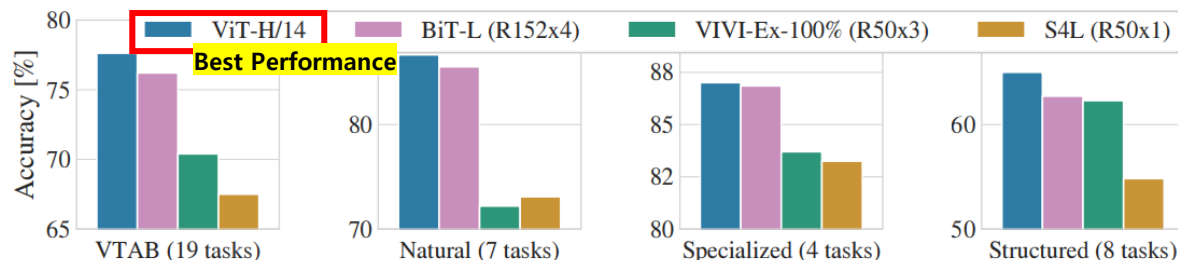
- **Experiment:**

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



벤치마크 된 lager 데이터셋으로 사전학습한 ViT와 SOTA 모델들의 데이터셋 별 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 ± 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

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 단점**을 가진 것 같다.