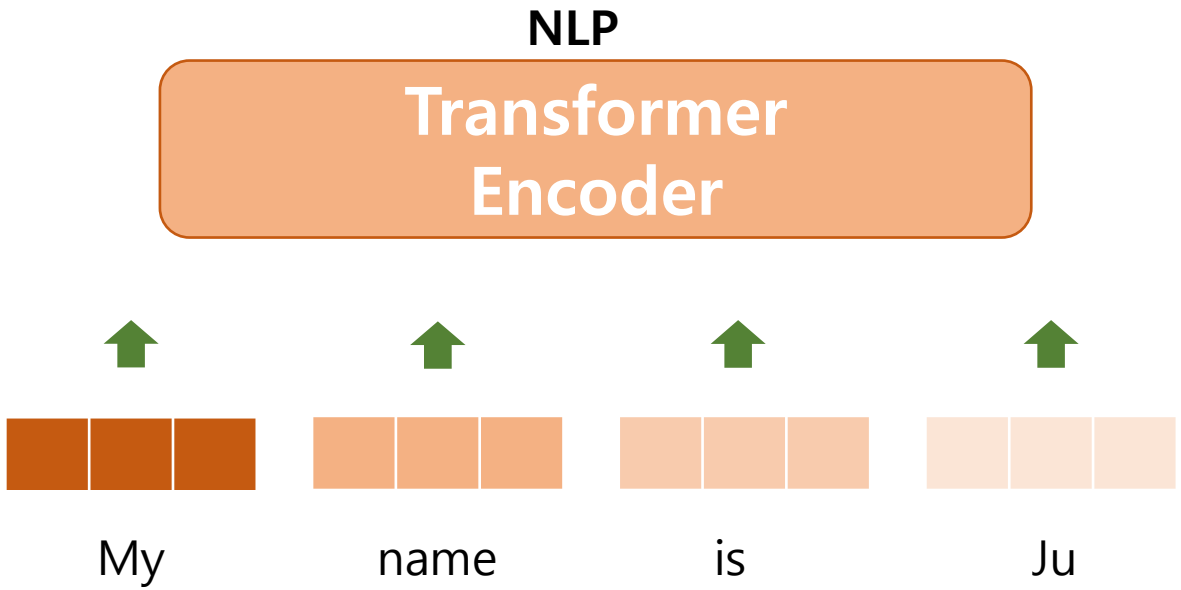
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al.

Google Research, Brain Team

Introduction

Vision Task (classification) 에 Transformer를 사용해보자
- transformer 를 직접적으로 이미지에 적용



Introduction

NLP

Transformer Encoder

My name is Ju

VISION

Transformer Encoder

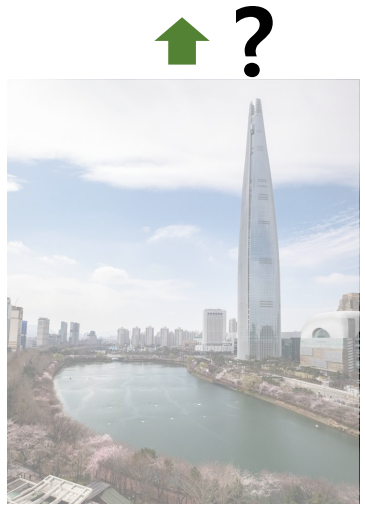
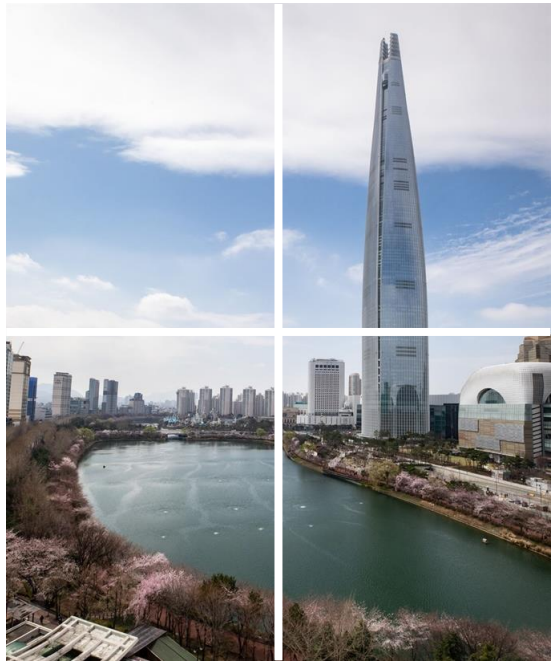


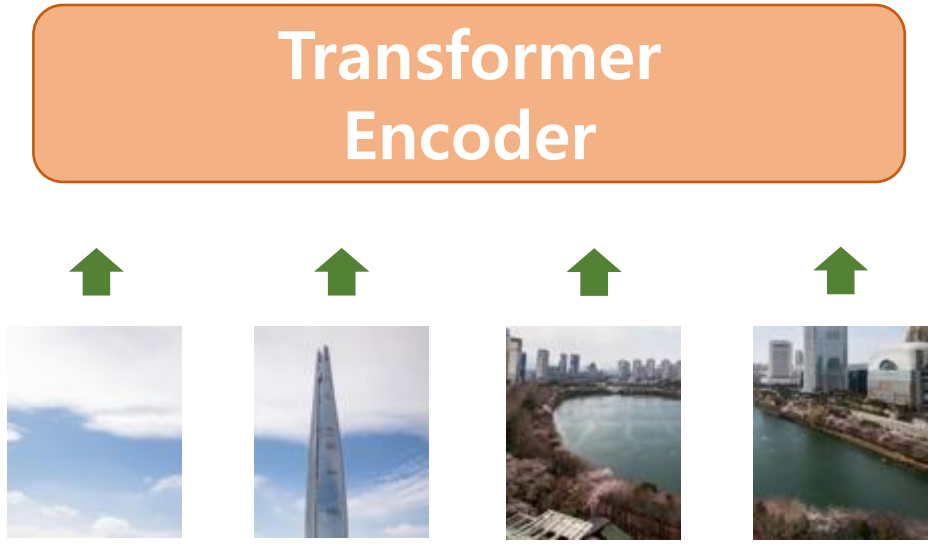
Image feature 를 어떻게 transformer에 적용 시켜야 하나

Introduction

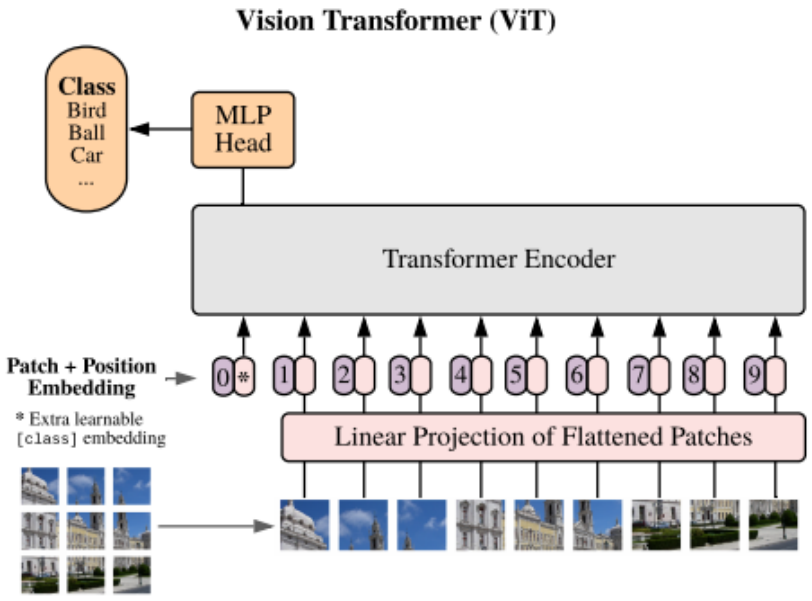
Image to Patch

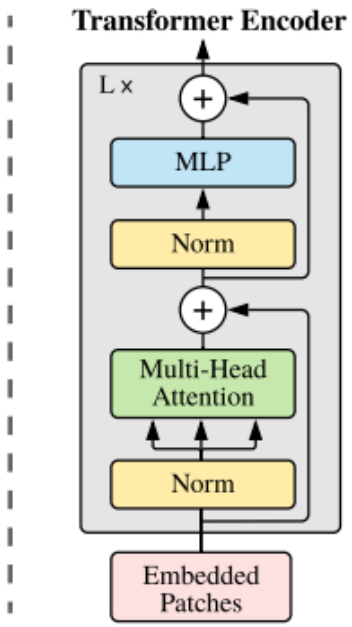


ViT



Method



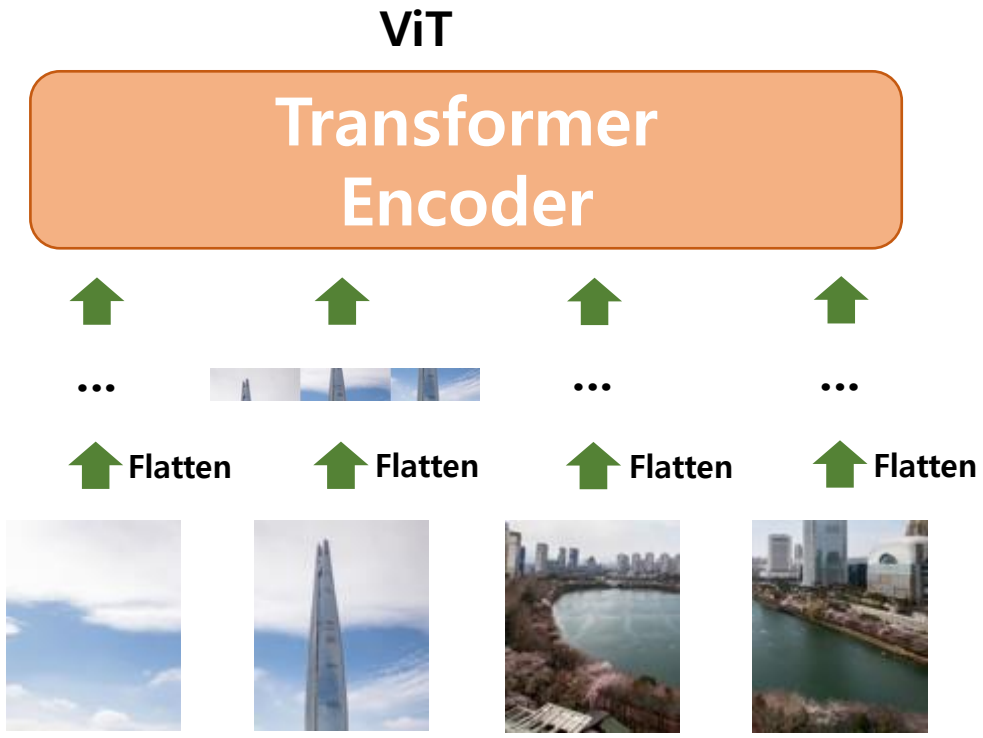


[INPUT]

일반적인 Transformer의 입력
- Token embedding 에대한 1차원의 시퀀스

이미지의 경우 flatten된 2차원의 패치의 시퀀스
- H x W x C → N x (P^2 x C) 로 변환
- (H, W) : 원본 이미지의 크기
- C : 채널 개수
- (P, P)는 이미지 패치의 크기
- N = HW/P^2 = 패치의 개수

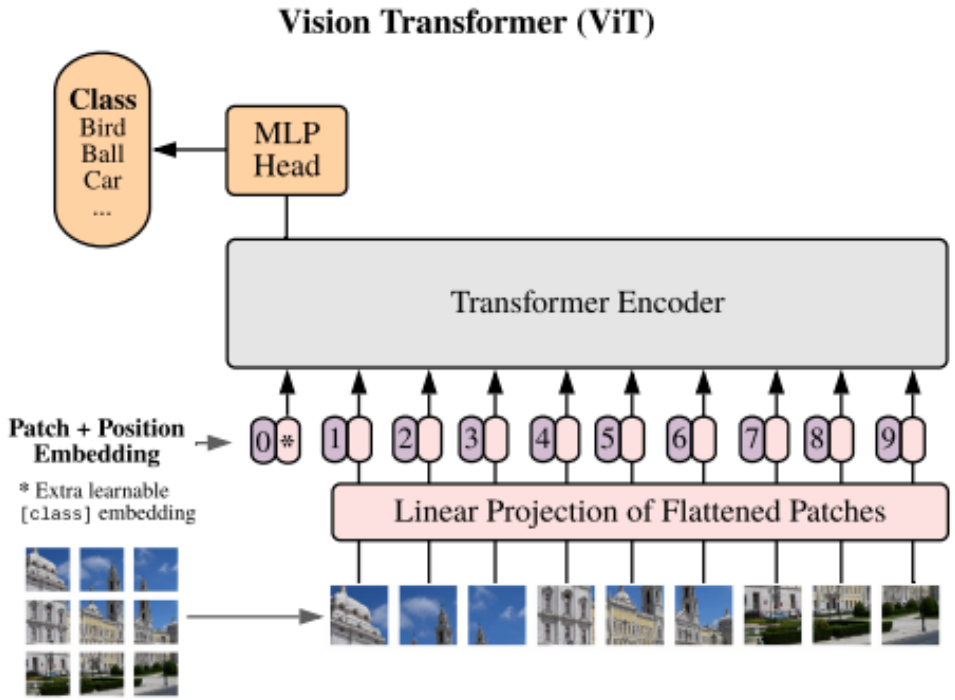
Method



[INPUT]
- 이미지 패치는 펼친 다음 D차원 벡터로 linear projection

[Embedding]
- 학습 가능한 1차원의 임베딩을 사용
- 2차원 정보를 유지하는 위치 임베딩도 유의미한 성능향상 X

Method



[INPUT]

- 임베딩 된 패치의 시퀀스에 z0 = x_class 임베딩을 추가 (BERT의 [CLS]토큰과 유사)
- 인코더 아웃풋은 이미지 representation으로 해석하여 분류

[Formulation]

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

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

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

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

Experiments

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.

Base 와 Large 모델의 경우 BERT 와 동일

(ViT-H/16 : 16*16의 패치 사이즈를 사용하는 ViT-Huge 모델)

Experiments

	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

- ViT-Huge 모델로 학습시킨 경우 SOTA의 성능을 보임
- ViT 의 Cost가 더 저렴함

- Metric : Top-1 Accuracy

Experiments

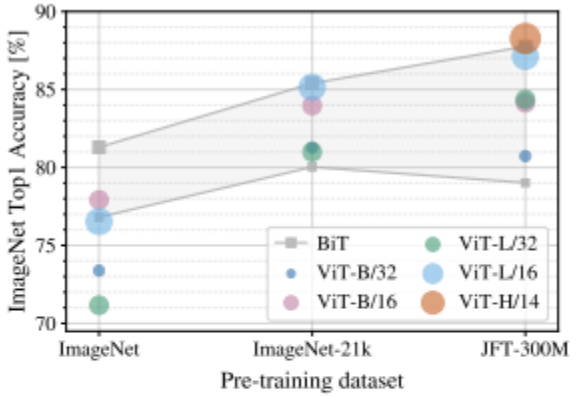


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

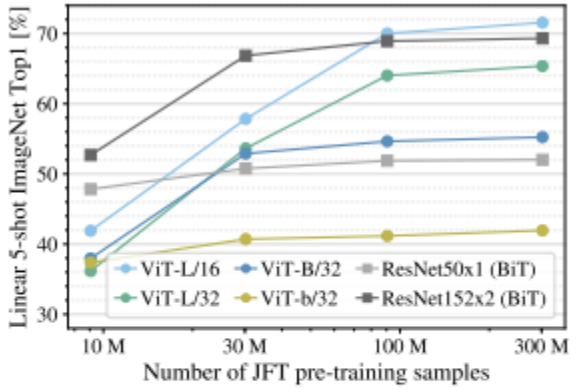


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Pre-training 한 데이터가 많아질수록 ViT의 성능 향상

Experiments

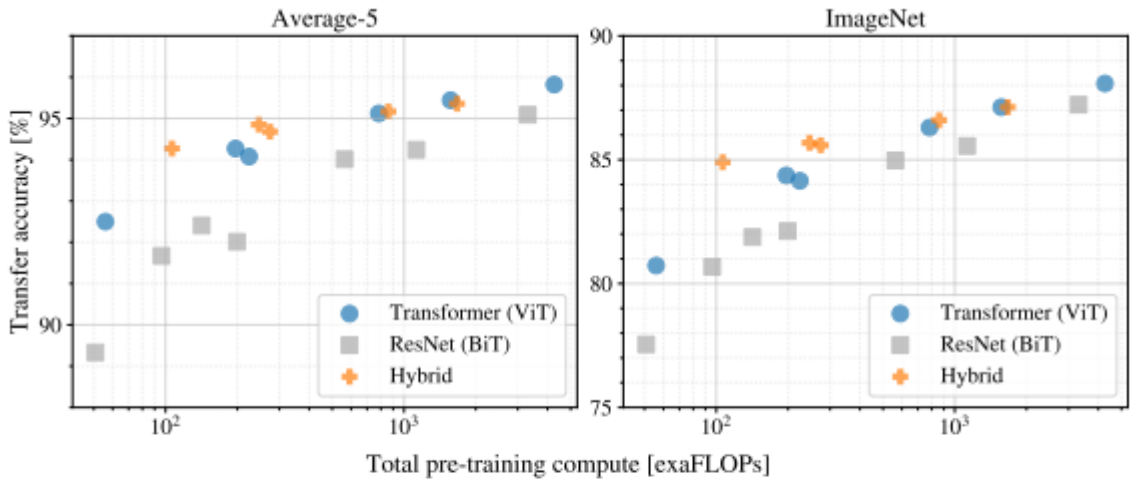


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

작은 모델 사이즈에서는 Hybrid 모델이 ViT보다 성능이 좋으나
큰 모델로 갈수록 격차가 사라진다
(Hybrid 모델 : CNN의 feature map으로부터 patch 생성)

Experiments

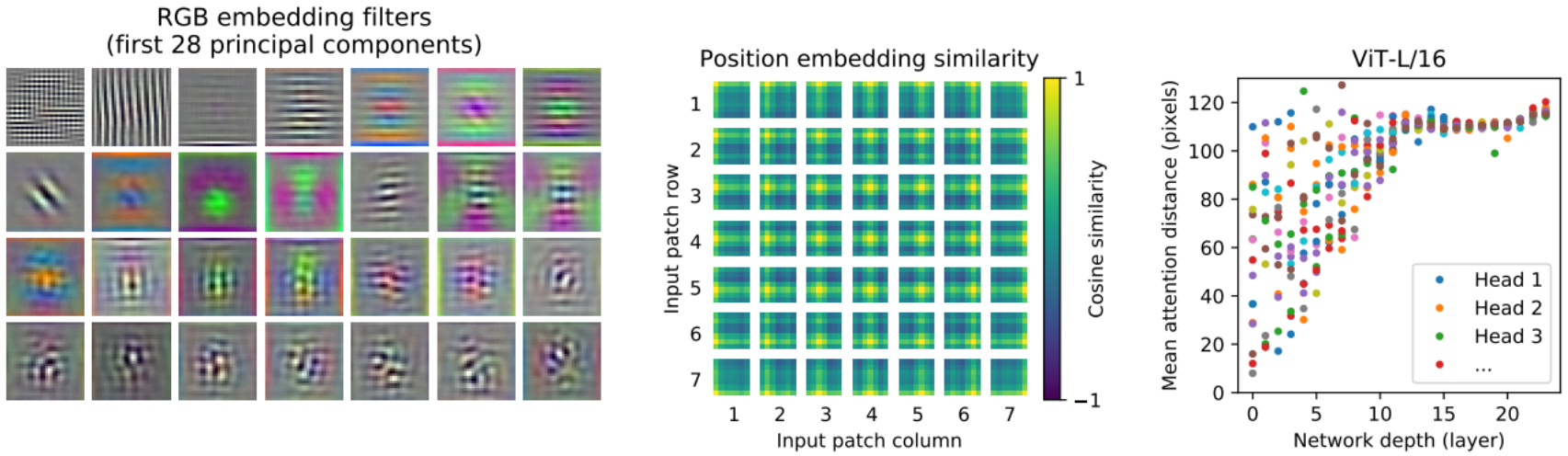
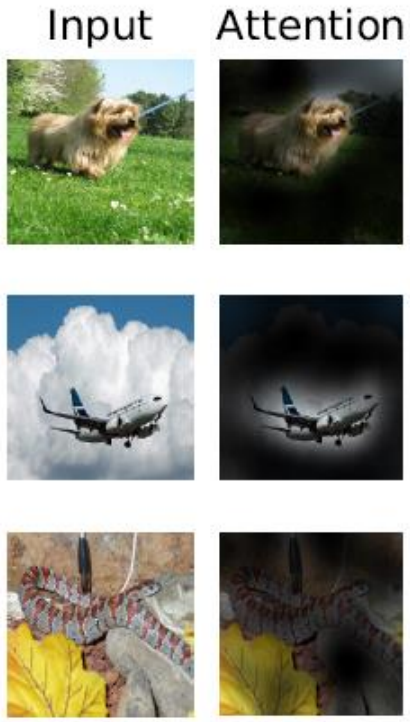


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

Experiments



ViT 가 분류에 있어서 의미 있는
영역을 찾는 것을 확인

Conclusions

- 큰 Dataset 에 pre-trained 되면 성능이 매우 우수
  (현 ImageNet Dataset SOTA)

- pre-training cost가 상대적으로 저렴하면서도 SOTA 성능 가능