An Image is Worth 16 X 16 Words: Transformers for image recognition at scale

발표자 : 윤태영(헬스케어IT공학과 석사 과정)

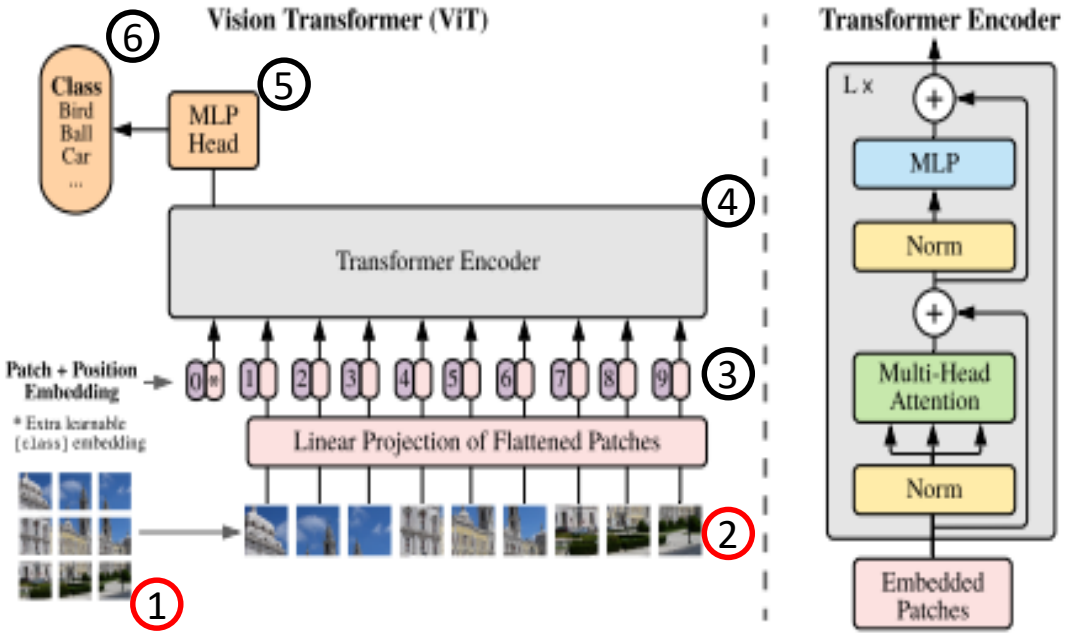
Introduction

➢ Vision transformer (ViT)

- NLP에서 사용되는 standard Transformer(self-attention-based architecture)를 이미지에 적용하여 이미지 분류를 제안.
- NLP의 transformer scaling에 영감을 받아, 이미지를 패치로 분할한 후, 이를 NLP의 단어로 취급하고 각 패치의 linear embedding을 순서대로 input으로 넣어 지도학습.
- ViT를 ImageNet-1k와 같은 적은 데이터를 학습했을 때 ResNet보다 낮은 정확도를 보임.
- 반면, ImageNet-21k와 같은 큰 데이터를 pre-train을 하고 다른 task로 transfer learning을 했을 경우 높은 정확도를 도출함.

Method

➢ ViT model



Step1 & 2

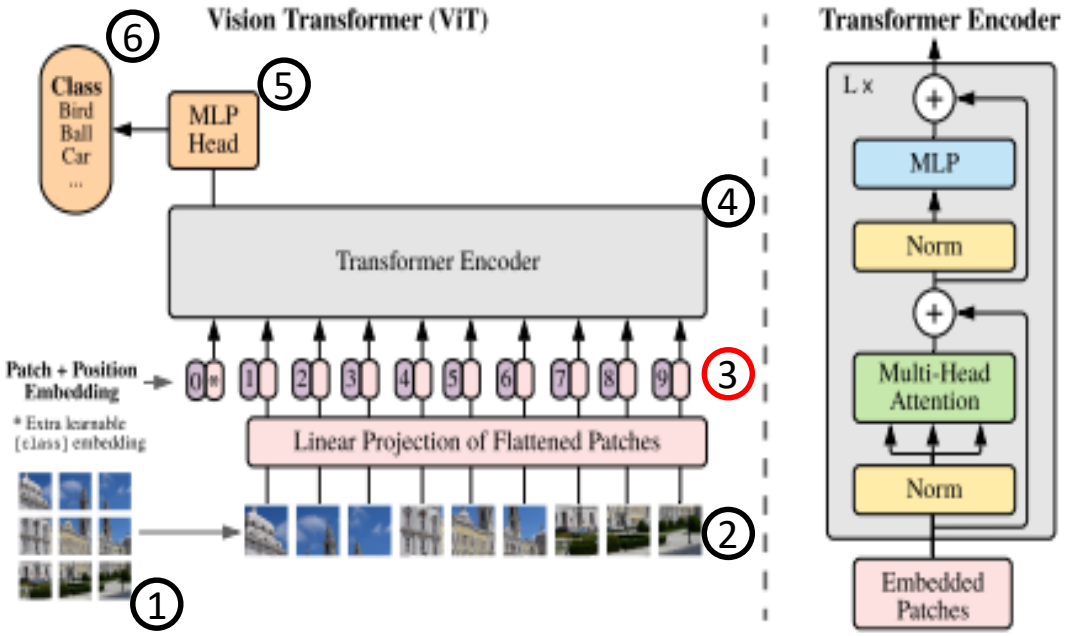
- **reshape the image into a sequence of flattened 2D patches**

$$X \in R^{H \times W \times C}$$
$$X^p \in R^{N \times (P^2 . C)}$$
$$N = HW/P^2$$

H,W = image
C = number of channel
(P,P) = each image path
N = resulting number of patches

Method

➢ ViT model



Step3

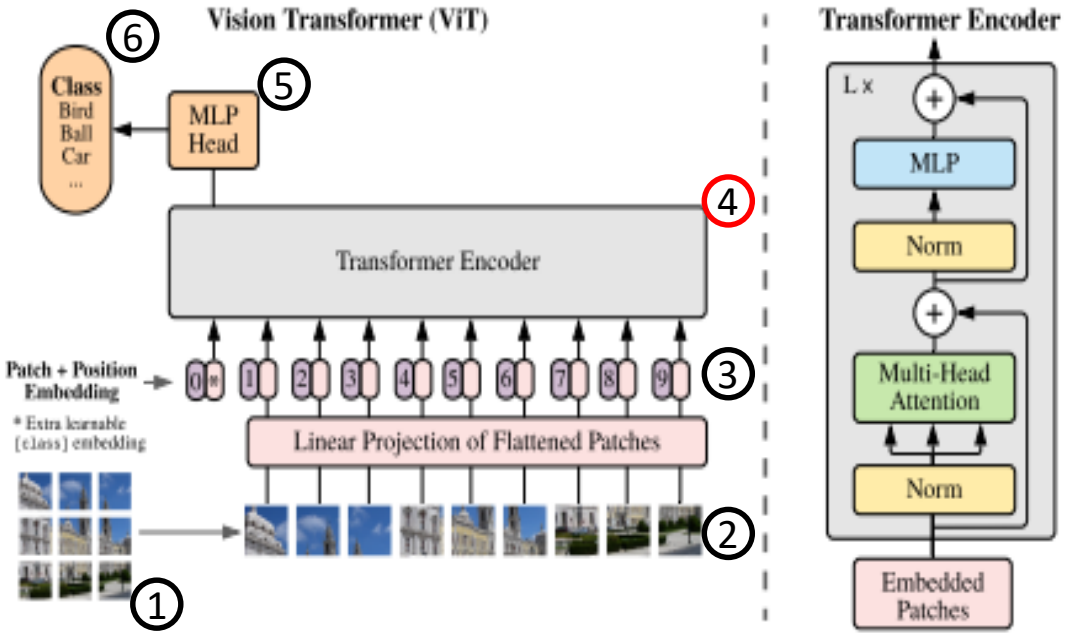
- **Trainable linear projection을 통해 X^p의 각 패치를 flatten한 패치 임베딩(1)**
- **Bert의 클래스 토큰과 유사한 learnable class임베딩(2)**
- **Class임베딩(2) 와 패치 임베딩(1)에 position임베딩 더함.**

Four different ways

- Providing no positional information : Consider input as a bag of patches
- 1-D positional embedding : Consider inputs as a sequence of patches in the raster order
- 2-D positional embedding : Consider input as a grid of patches in two dimensions
- Relactive positional embedding : Consider relative distance between patches to encode the spatial information as instead of their absolute position

Method

➢ ViT model

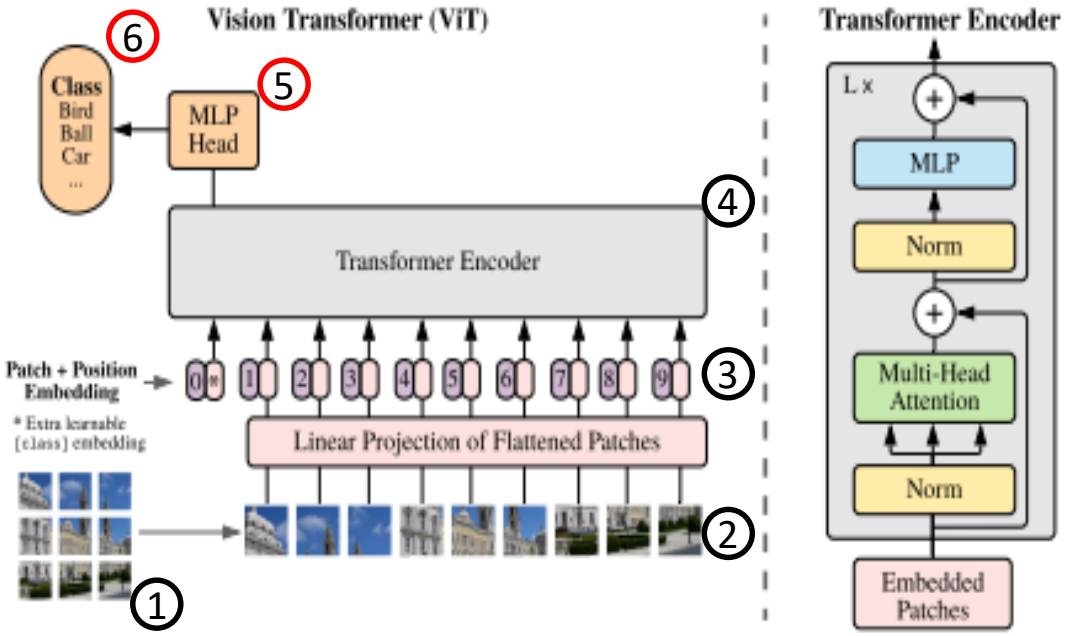


Step4

- Step3에서의 임베딩을 vanilla Transformer encoder에 넣음.
- Transformer encoder는 2개의 layer norm을 각각 매 블록이전에 넣고, 2개의 residual connection을 각 각 매 블록 후에 넣음.

Method

➢ ViT model



Step5, 6

- **Transformer Encoder(step4)의 output인 image representation을 입력으로 넣어 이미지의 클래스를 분류함.**

Method

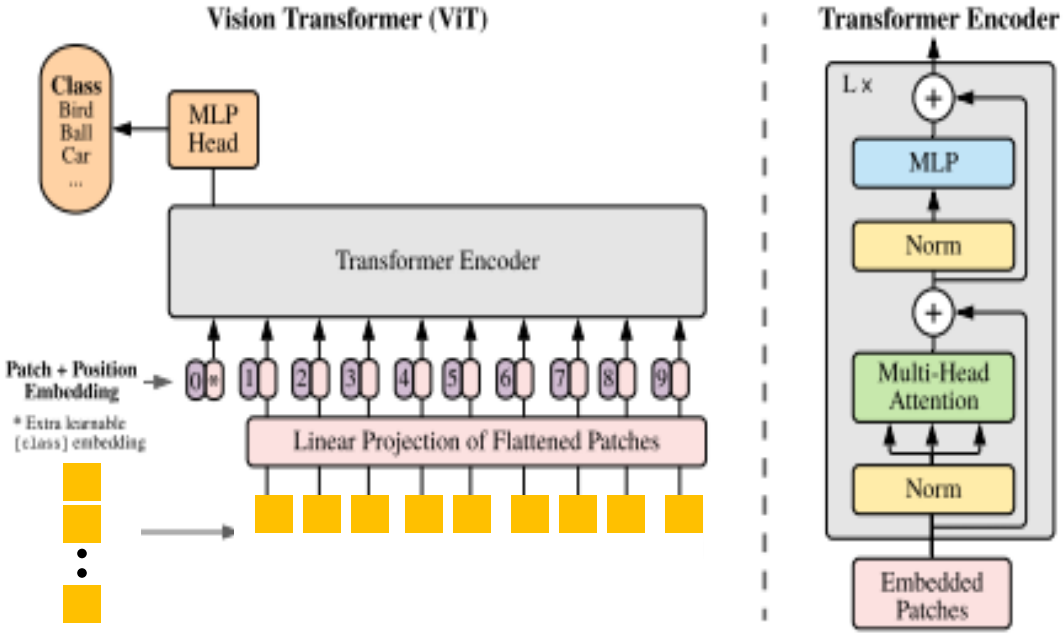
➢ Inductive bias

- Inductive bias는 학습과정에서 보지 못한 데이터의 적절한 귀납적 추론이 가능하도록 하기 위해 가지고 있는 가정들의 집합.
- Transformer는 CNN & RNN보다 inductive bias가 낮음.
- ViT의 경우 MLP는 locality와 translation equivariance, MSA는 global
  그러므로 CNN보다 image-specific inductive bias이 낮다.
- ViT에서는 두가지 방법을 사용하여 inductive bias을 주입
  1) Patch extraction : cutting the image into patches
  2) Resolution adjustment(해상도 조정) : adjusting the position embeddings for images of different resolution

Method

➢ Hybrid Architecture

- Raw image patches가 아닌 CNN에서 추출된 feature maps을 input sequence로 사용 가능함.
- Feature map의 패치들은 1 X 1으로 간단하게 차원을 flatten하고 linear projection을 하면 됨.



Method

➢ Fine-tuning & Higher Resolution

- ViT는 large data를 사용하여 pre-training을 하고 downstream task로 fine-tuning 진행.
  - ✓ Pre-trained prediction head를 제거
  - ✓ Feedforward layer에 zero-initialized D X k 삽입(k = the number of stream classes)
  - ✓ 때때로, fine-tuning까지 진행을 했을때 pre-train보다 좋은 결과를 보임.

- When feeding images of higher resolution
  - ✓ Patch size는 동일 하기 때문에 sequence length가 증가.
  - ✓ Pre-train의 위치 임베딩은 더 이상 의미가 없기 때문에, raw image의 위치에 따라 pre-trained 위치 임베딩을 2D보간을 진행

Experiments

➢ Datasets

- ViT는 아래 표와 같이 각각 다른 3개의 데이터셋을 기반으로 pre-train
- Several benchmark tasks :

  ReaL labes, CIFAR-10/100, Oxford-IIIT Pets

  Oxford Flowers-102, 19-task VTAB

Pre-trained Dataset	classes	Images
ImageNet-1k	1k	1.3M
ImageNet-21k	21k	14M
JFT	18k	303M(high resolution)

Experiments

➢ Model Variants

- BERT에서 사용되는 구성인 Base, large에 Huge추가적으로 진행
- ResNet(CNN)같은 경우 Batch Norm을 Group Norm을 사용하고, 표준화된 convolution을 사용("ResNet(BiT)")

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

➢ Train

- $\beta_1 = 0.9,\ \beta_2 = 0.999$ Adam, batch size = 4096, weight decay = 0.1

➢ Fine-tune

- Momentum의 SGD, batch size = 512
- Higher resolution fine-tune : ViT-L/16 → 512, ViT-H/14 → 518

Experiments

➢ Comparison to State Of The Art

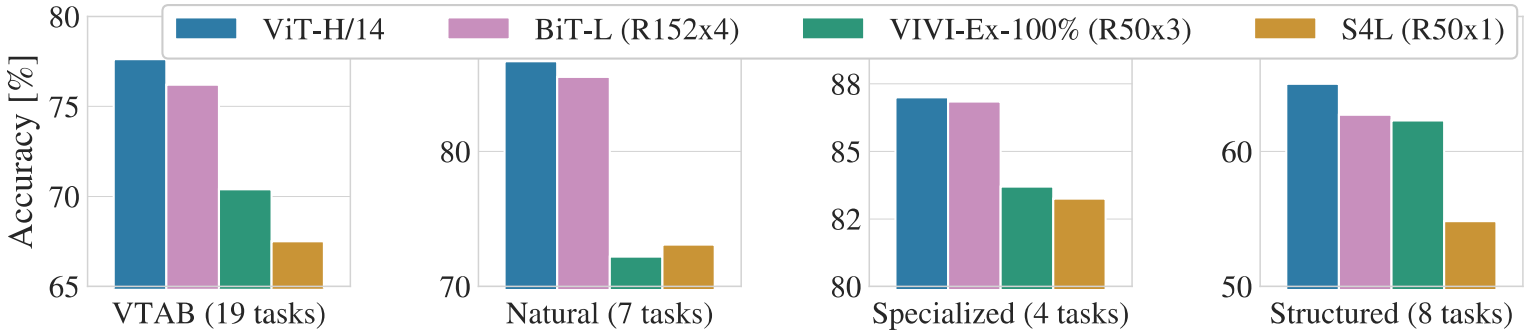
- **Comparison points**
  1) **Big transfer(BiT) : 대용량 ResNets의 supervised transfer learning**
  2) **Noisy student : label을 제거한 ImageNet과 JFT-300M으로 semi supervised learning을 한 대용량 EfficientNet**

	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

Experiments

➢ Comparison to State Of The Art

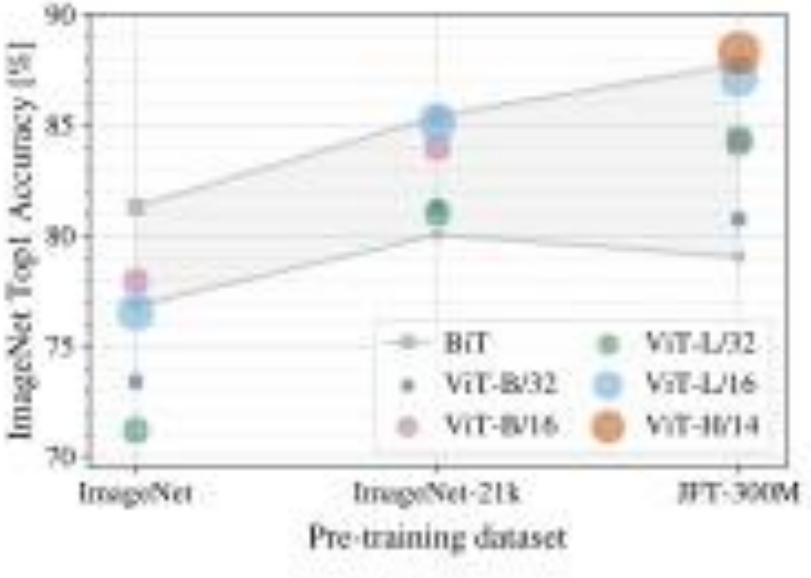
- 19-task VTAB classification suite를 3개의 그룹으로 나눠 실험 진행.
  - ✓ Natural : Pets, CIFAR, etc.
  - ✓ Specialized : medical and satellite imagery
  - ✓ Structured – tasks that require geometric understanding like localization

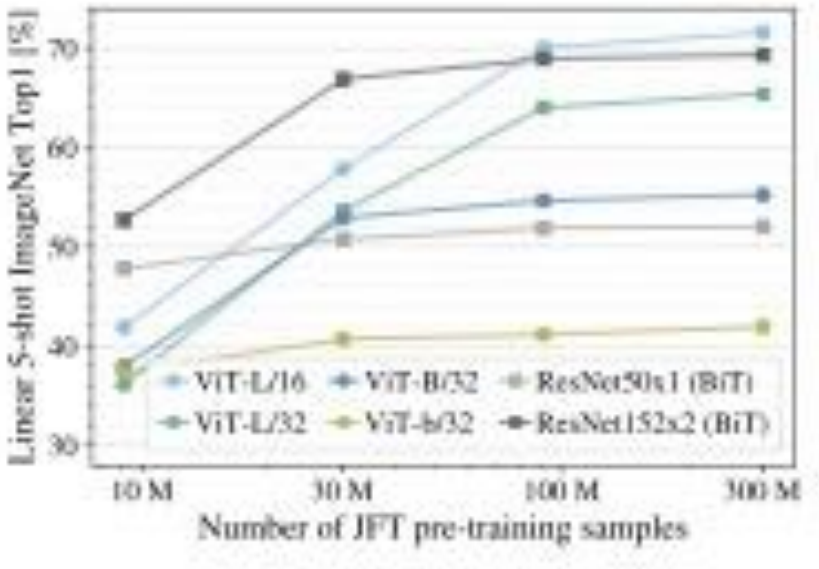


Experiments

➢ Pre-training data requirements

- Dataset의 크기가 증가함에 따라 BiT보다 ViT 성능이 좋고 large ViT 모델의 효과가 크게 나타남.
- pre-train에 사용한 ImageNet 사이즈에 따른 결과를 보면 ResNet은 사이즈가 증가함에 변화를 보이지 않지만 ViT모델의 경우 성능이 증가

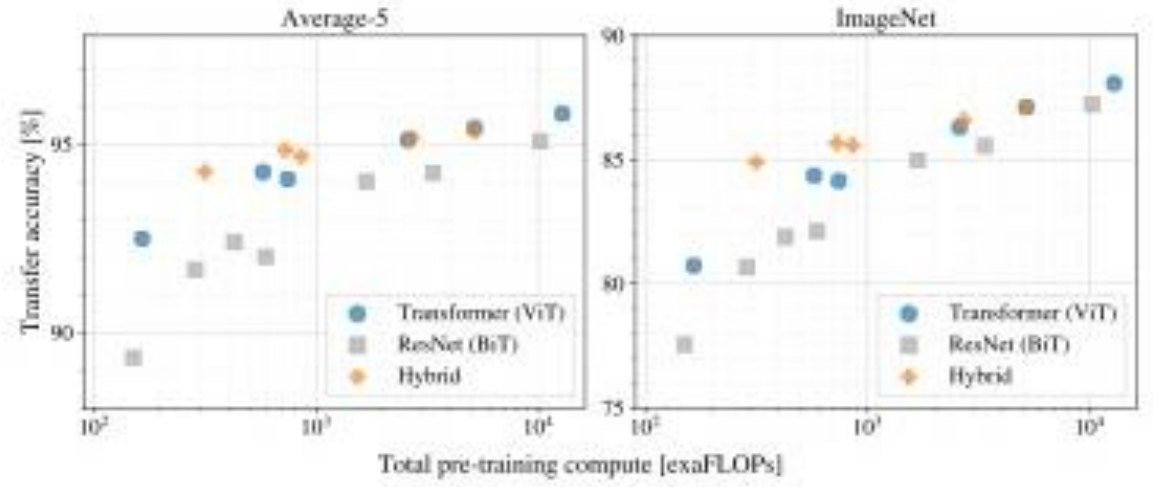




Experiments

➢ Scaling study

- ViT는 performance/compute trade-off에서 ResNet 압도
- ViT는 동일 성능 대비 2-4배 적은 computing cost
- 하이브리드 모델의 경우 적은 컴퓨팅에서는 성능이 우수하지만 더큰 모델에서는 그 차이가 사라짐



Experiments

➢ Inspecting Vision Transformer

- (left) : 학습된 embedding filter의 주요한 구성요소. 각 패치내에서 미세 구조의 저차원 표현을 위한 기본함수와 비슷함.
- (center) : 가까운 패치가 더 유사한 위치 임베딩을 갖는 경향이 있음을 보여줌.
- (right) : layer별 평균 attention distance를 보면, 초반 layer에서도 이미지 전체의 정보를 사용함을 보여줌.