ELLab

SUNG KYUN KWAN UNIVERSITY

Vision Transformer

2025.11.28

Introduction

ViT

- Transformer는 NLP에서 대규모 사전학습을 기반으로 표준 모델이 되었지만 Vision 분야에서는 CNN이 여전히 주류임
  - 기존의 self-attention 기반 비전 모델은 하드웨어 비효율 때문에 대규모 확장이 어려웠음

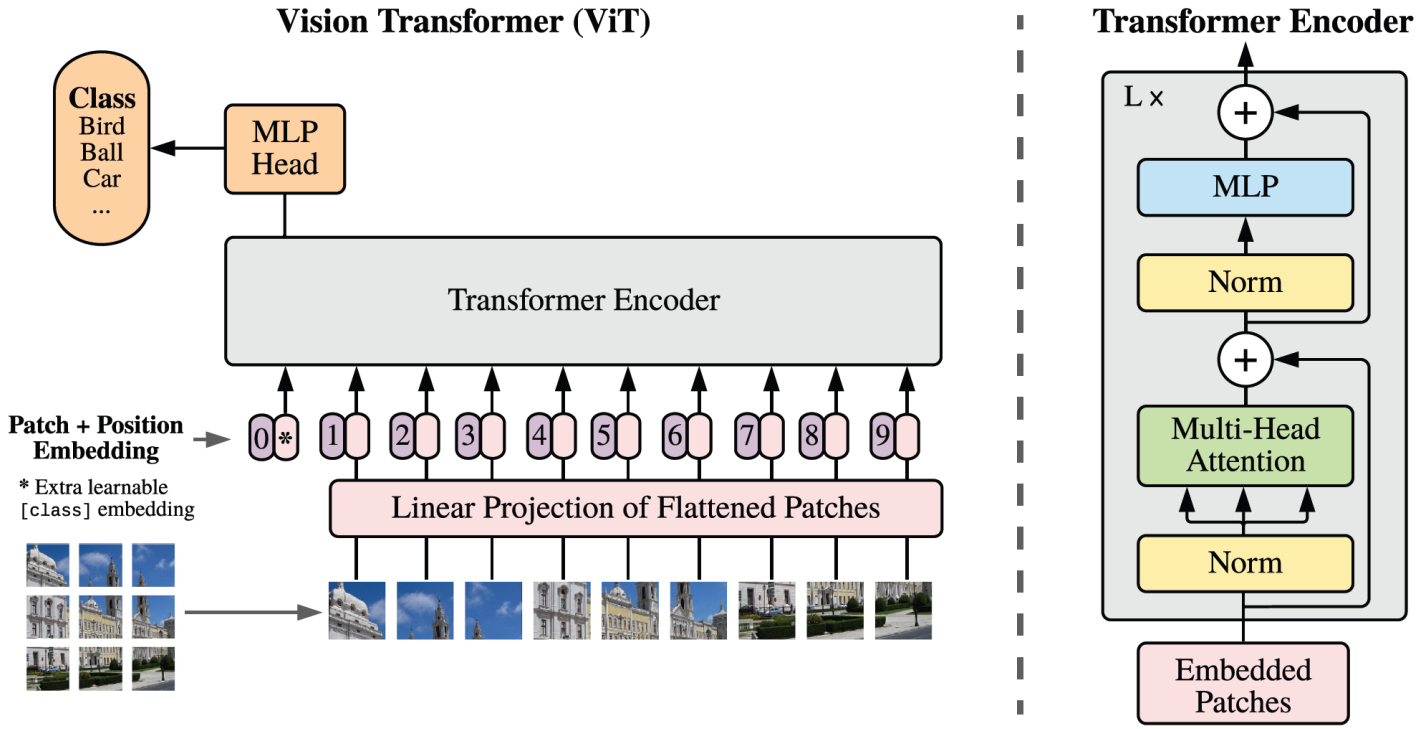
- ViT는 이미지를 패치 단위 토큰으로 보고 Transformer를 그대로 적용한 단순한 구조임
  - 그러나 중간 규모 데이터(ImageNet)만으로 학습하면 CNN보다 성능이 안좋음

- 매우 큰 데이터셋(14M~300M)으로 Pretrain하면 성능이 급격히 개선됨
  - 데이터 규모가 커지면 CNN의 inductive bias 부족을 충분히 보완할 수 있기때문

- 대규모 사전학습 ViT는 다양한 벤치마크에서 CNN SOTA를 능가함
  - ImageNet 88.55%, CIFAR-100 94.55%, VTAB 77.63% 등 최고 수준의 성능 기록

Vision Transformer

Architecture



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

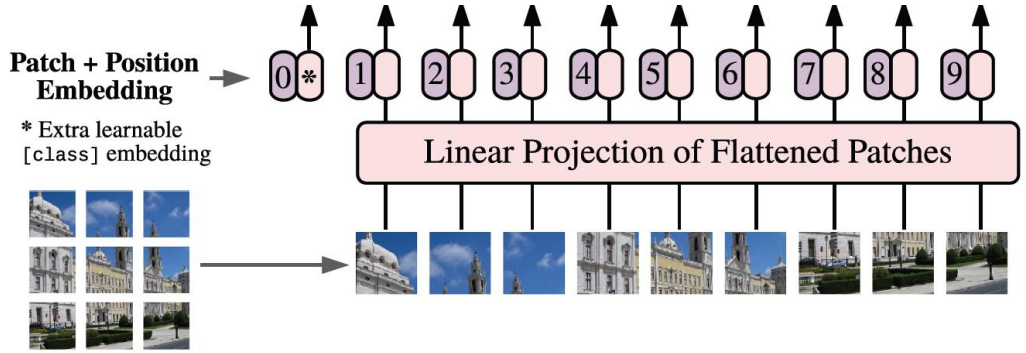
$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

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

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

Vision Transformer

Embedding



Reshape $x \in R^{H \times W \times C} \rightarrow x \in R^{N \times (P^2 \cdot C)}$

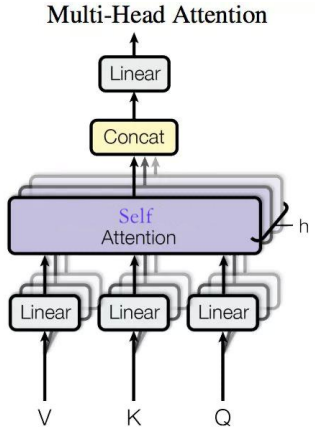
- 입력 이미지 $x \in R^{H \times W \times C}$ 를 $P \times P$크기의 패치로 분할하여 시퀀스로 변환함
- 각 패치는 flatten되어 $x_p \in R^{(P^2 \cdot C)}, N = \frac{HW}{P^2}$형태의 벡터가 됨
- 모든 패치는 학습 가능한 선형 projection을 통해 공통 임베딩 차원 D로 변환됨
- 전체 패치 개수는 $N = \frac{HW}{P^2}$ 이며, 이는 Transformer의 입력 시퀀스 길이가 됨
- 이미지 전체를 대표하는 class token x_{class}을 시퀀스 맨 앞에 추가함
- self-attention에 위치 정보를 제공하기 위해 학습 가능한 positional embedding을 더함
- 최종적으로 생성된 임베딩 시퀀스가 Transformer encoder의 입력으로 사용됨

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

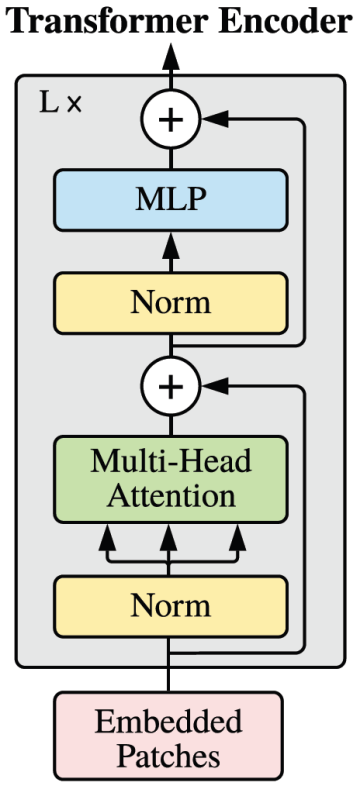
Vision Transformer

Transformer Encoder

- 계산된 attention 가중치 AA 를 Value에 곱해 weighted sum 수행
  - ➢ $SA(z) = A\,v$



- k개의 서로 다른 head에서 self-attention을 병렬로 수행한 뒤 결과를 concat하고 U_{msa}로 선형 변환해 원래 차원 D로 돌려놓음
  - ➢ $MSA(z) = [SA_1(z);;SA_2(z);;\ldots;;SA_k(z)], U_{msa}$



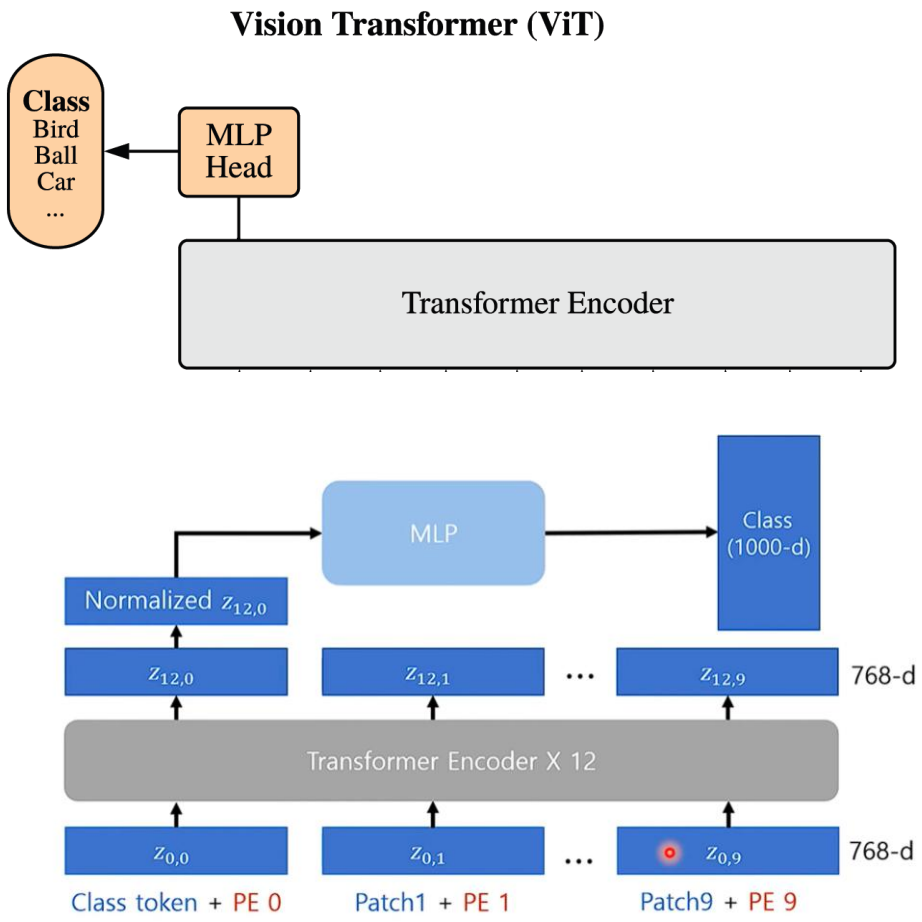
- MLP
  - Self-Attention을 거친 중간 표현 z'_l을 다시 LayerNorm으로 정규화하고 MLP에 통과시켜 비선형 변환을 수행함 → 패치간 정보를 섞어줌
- 이 과정을 L번 반복

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

Vision Transformer

Classification

- Transformer Encoder의 출력 토큰 중 CLS 토큰(z_L^0)이 이미지 전체를 대표하는 feature로 사용됨
- 최종 feature에 LayerNorm을 한 번 더 적용해 안정적인 표현을 만든 뒤 MLP Head(Linear Layer) 를 통해 클래스 확률로 매핑함
  - Appendix D.3에 single hidden layer에 tanh가 쓰였다고 나옴



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

Vision Transformer

Properties

Inductive bias

- CNN은 locality, translation equivariance 등 강한 이미지 inductive bias를 가짐
  - 이웃 픽셀 구조가 레이어 전체에 자연스럽게 반영됨
- ViT는 초기 패치 분할 외에는 이미지 구조 사용 X
  - Self-attention은 전역 비교 → 2D 위치 정보는 학습을 통해 스스로 학습해야함

➢ 따라서 작은 데이터에서는 불리하지만, 대규모 학습 시 더 강한 표현력 발휘

Hybrid Architecture

- 원본 이미지를 자르는 대신, CNN의 feature map을 패치처럼 사용 가능

➢ CNN의 inductive bias + ViT의 전역 표현력을 결합한 hybrid 모델

Vision Transformer

Properties

Fine tunning & Higher resolution

- ViT는 대규모 데이터로 pre-train 후 작은 downstream task에 fine-tuning함

- Fine-tuning 시 해상도를 pre-training보다 크게 하면 성능 향상
  - patch size는 그대로 → patch 수 증가 → 더 긴 시퀀스를 Transformer가 처리

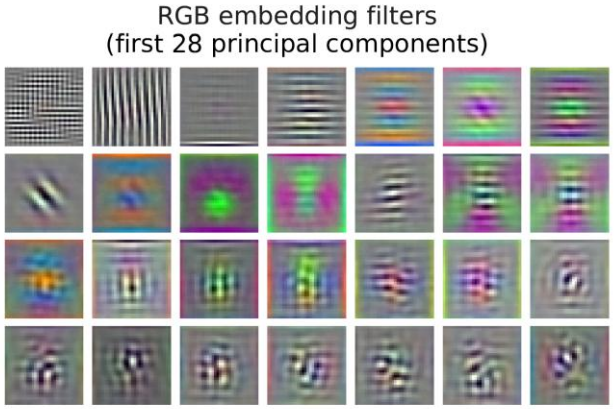
- Transformer는 가변 길이 시퀀스를 자연스럽게 처리 가능

- 하지만 pre-trained 위치 임베딩은 원래 해상도에 맞춰져 있어 의미가 달라짐

➢ 위치 임베딩을 fine-tuning 해상도에 맞게 2D interpolation 함

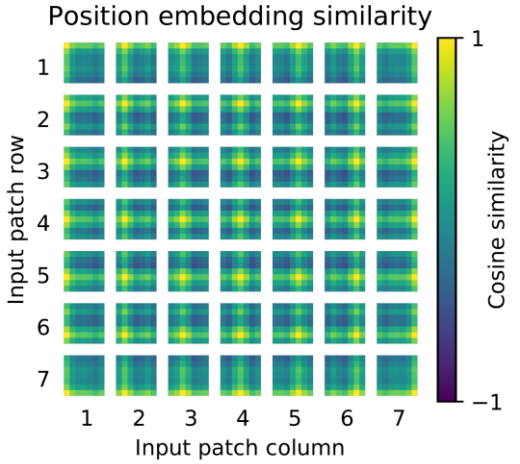
Experiment

Inspecting ViT



Patch Embedding

- ViT의 첫 layer는 patch를 flatten하여 Linear projection함
- PCA로 시각화하하면 projection filter들이 패치 내부의 미세 구조를 잡아내는 기저 함수처럼 보임



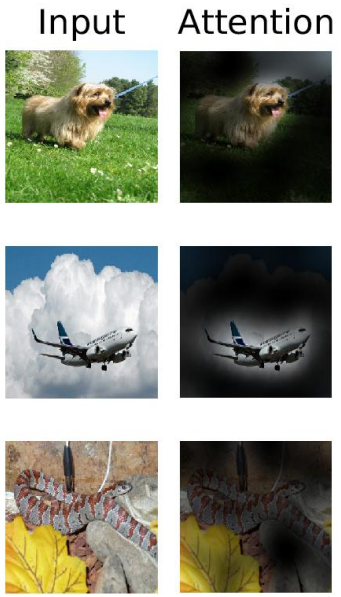
Position Embedding

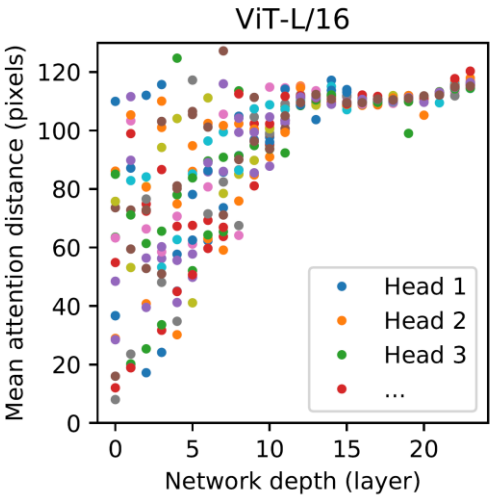
- 가까운 패치는 비슷한 embedding을 가짐
- row/column 방향 구조도 자연스럽게 학습됨

➢ 즉, ViT는 2D topology를 스스로 학습할 수 있음

Experiment

Inspecting ViT





Self-Attention

- 초기 layer부터 이미 전체 이미지로 넓게 attention을 펼치는 head가 존재함
  - ViT는 receptive field가 CNN보다 훨씬 빠르게 전역으로 확장됨
- 반면 일부 head는 low-level의 local attention을 유지
  - 이러한 local attention은 Hybrid 모델(CNN → ViT)에서 더 뚜렷함
- deeper layer로 갈수록 attention distance 증가하고 classification 토큰은 semantic region에 집중함

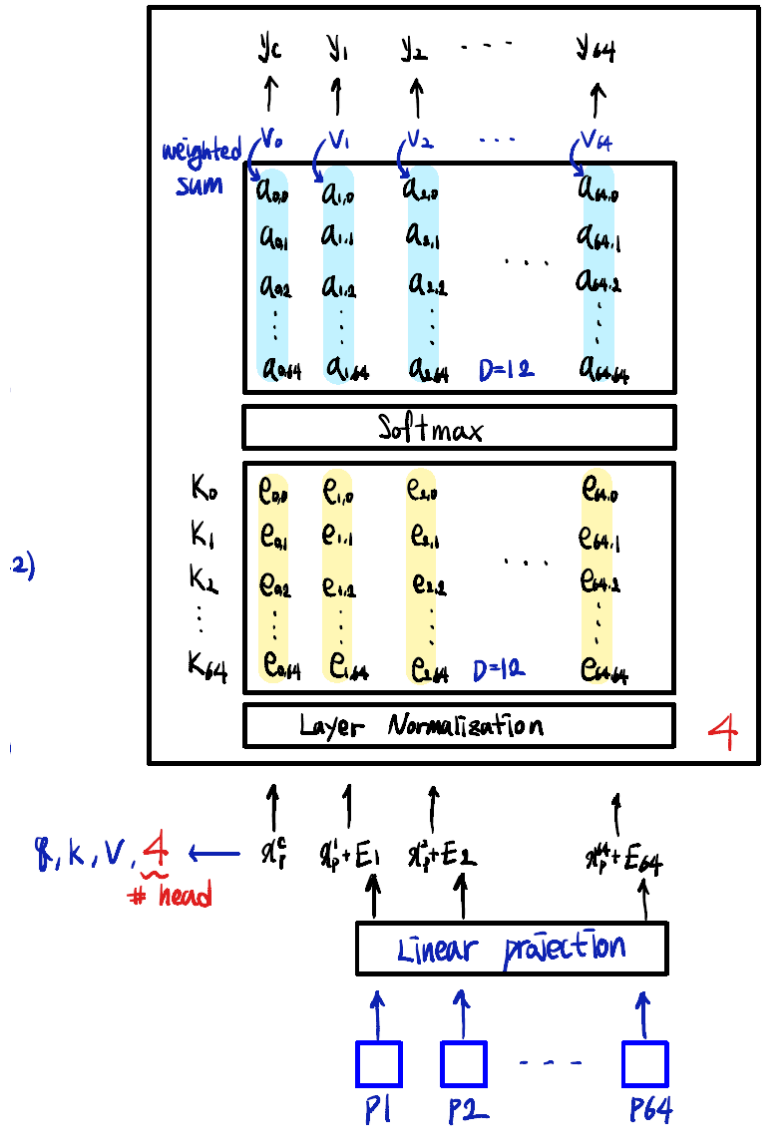
➢ ViT는 전역적+국소적 정보를 동시에 학습하는 구조이며 receptive field를 확장 가능함

My Code

Create ViT Model

- Patch Embedding
- Multihead Attention
- Residual
- Feedforward Block
- Transformer Encoder Block
- Transformer Encoder
- Classification Head
- ViT

https://colab.research.google.com/drive/1bdjcEwn8pwvLSOnfnwgeTVH9tA6nCkwy?usp=sharing



My Code

Create ViT Model

Dataset	score
CIFAR10	71.36%
CIFAR100	44.08%

ELLab