Vision Transformer(ViT)

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

Google Research, Brain Team
Oct. 2020 / ICLR 2021

210527 한지수 (그림 출처는 논문)

NLP 추세

- 요즘은 Transformer가 기본적인 모델! LSTM 시대는 지나감.
- Transformer 응용
  - BERT: Pre-train on a large text corpus → Fine-tune on a smaller task-specific dataset
  - GPT
  - Transformer & Vision
    - VideoBERT
    - iGPT
    - ViT 등등

Attention

- Decoder → Query / Encoder→ Key, Value
- Encoder, Decoder와의 상관관계를 바탕으로 추출

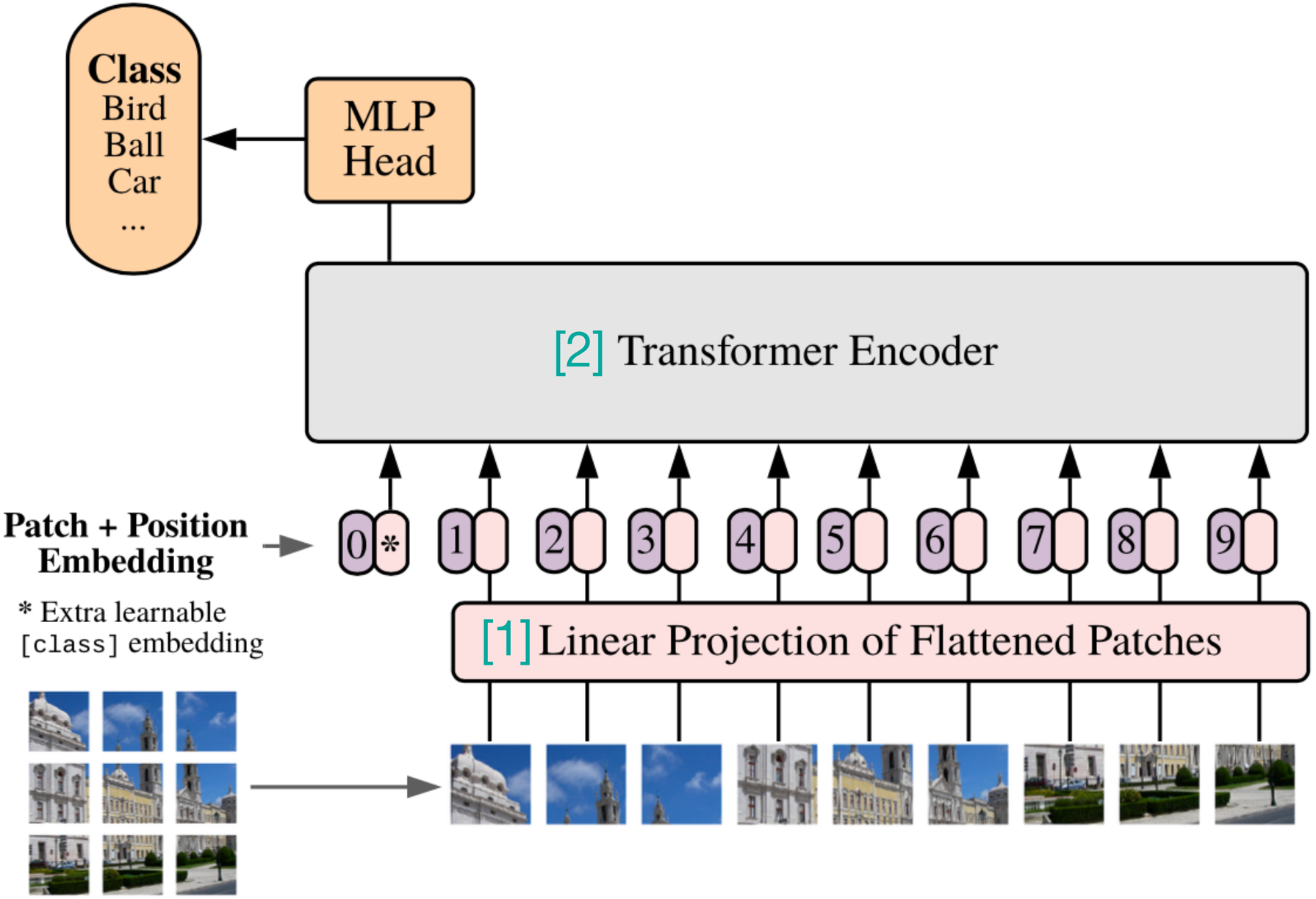
Transformers(Attention is all you need)

- 입력 데이터 → Query, Key, Value
- Decoder가 특정시점 단어를 출력할 때 Encoder 정보 중 연관성이 잇는 정보를 직접 선택(Self-Attention)
    - 데이터 내의 상관관계를 바탕으로 추출
- Inductive bias 가 거의 없어 많은 수의 데이터가 필요 → 후의 ViT 학습에 영향

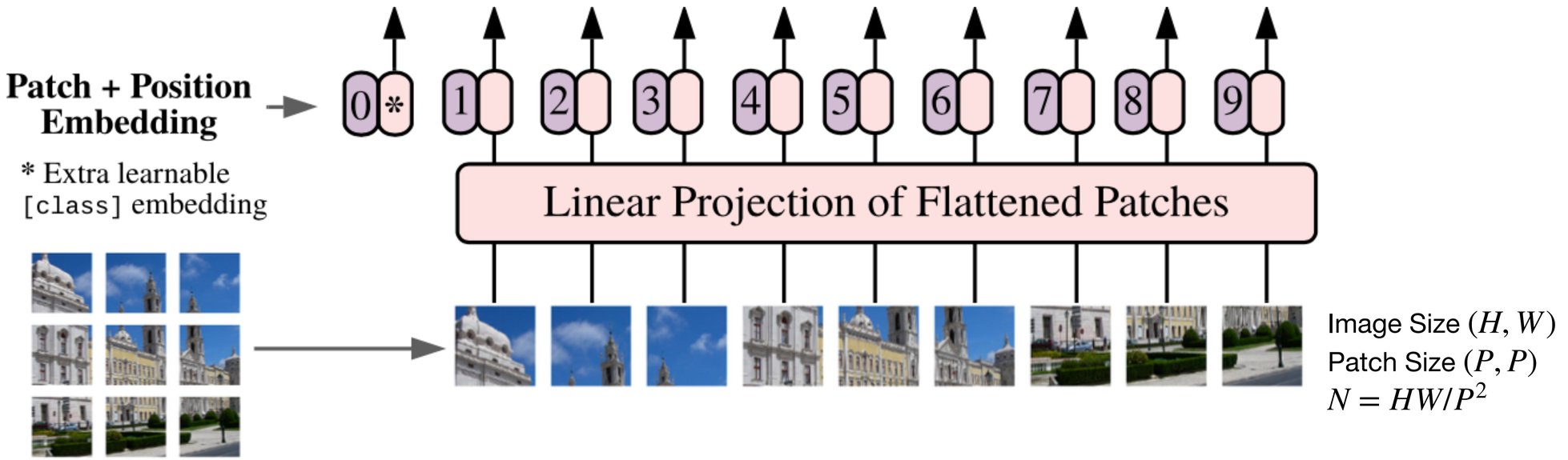
Transformers & Vision

- 하나의 Layer로 전체 이미지 정보 통합

Vision Transformer (ViT)



[1]



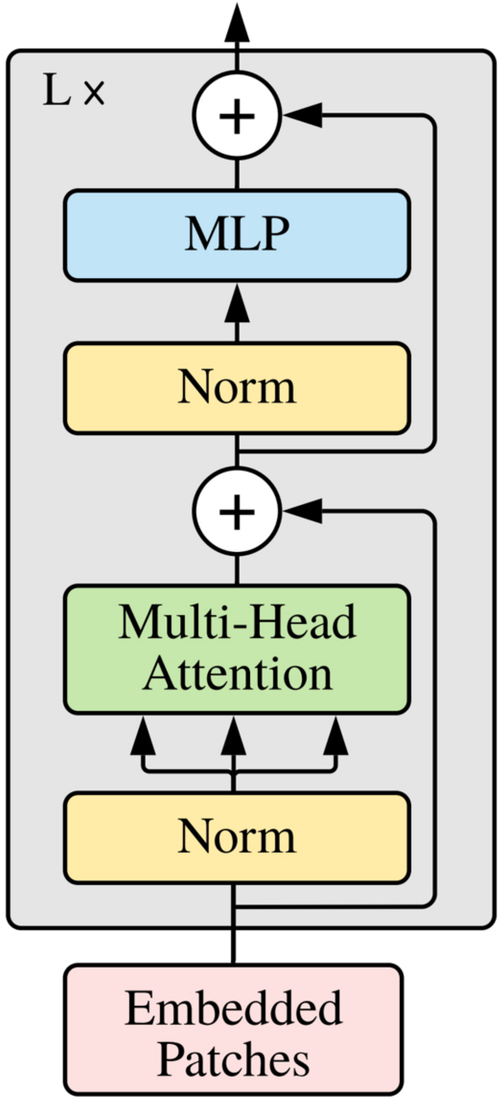
- Reshape the image $x \in \Re^{H \times W \times C}$ into a sequence of flattened 2D patches(tokens) $x_p \in \Re^{N \times (P^2 \times C)}$
- Eq 1. Embedded Patches

  $z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}$, N개의 patch, p번째 이미지, position embedding E_{pos}

  $E \in \Re^{(P^2 * C) \times D}, E_{pos} \in \Re^{(N+1) \times D}$
- Hybrid Architecture

  이미지를 패치로 분할하는 대신 ResNet으로 처리 가능

[2] Transformer Encoder



Eq 4. $y = LN(z_L^0)$ LN Layernorm

Eq 3. MLP

$$z_l = MLP(LN(z'_{l-1})) + z'_{l-1}, l = 1...L$$

LN Layernorm & Residual Connection
MLP two layers with a GELU non-linearity

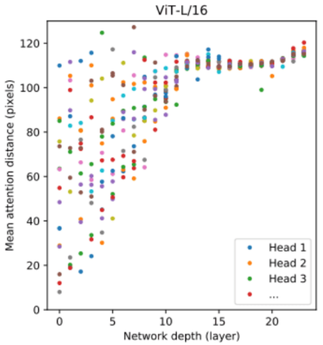
Eq 2. Multiheaded Self-Attention

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, l = 1...L$$

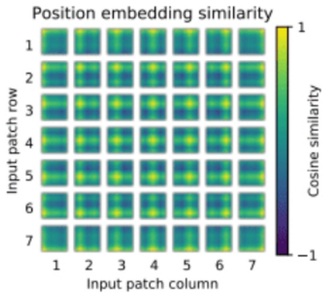
LN Layernorm & Residual Connection

Eq 1. Embedded Patches

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}$$



Self-Attention으로 전체 이미지에 대한 정보 통합 가능
/ 이미지 정보가 통합되는 이미지 공간의 평균 거리를 계산
/ CNN의 receptive field size와 유사



Position Embedding의 유사성으로 이미지내 거리를 학습함.
가까운 patch가 유사한 position embedding을 지닌다.

Etc.

- Fine-tuning and Higher resolution
  - Pre-train보다 높은 resolution으로 fine-tuning이 괜찮다.
- Pre-training data requirements
  - CNN 고유의 inductive bias를 고려할 수 있는 기능이 Transformer에 없어 많은 데이터를 요구
    → JFT-300M(3억 이미지, 10억 라벨)로 구글만이 해낼 수 있었다.
  - Large-scale 데이터 셋에서는 inductive bias를 능가한 높은 결론이 가능하다.
- Recognition 벤치마크에서 SOTA 달성

확장 | Data-efficient Image Transformer (DeiT)

- “Training data-efficient Image Transformers & distillation through attention” (Facebook)
  - 앞의 ViT에서의 데이터셋 구축 필요 없이, pre-training 필요 없다고 함
  - 요인 1) Data augmentation, Optimization, Regularization 강화
  - 요인 2) knowledge distillation - distillation token 추가

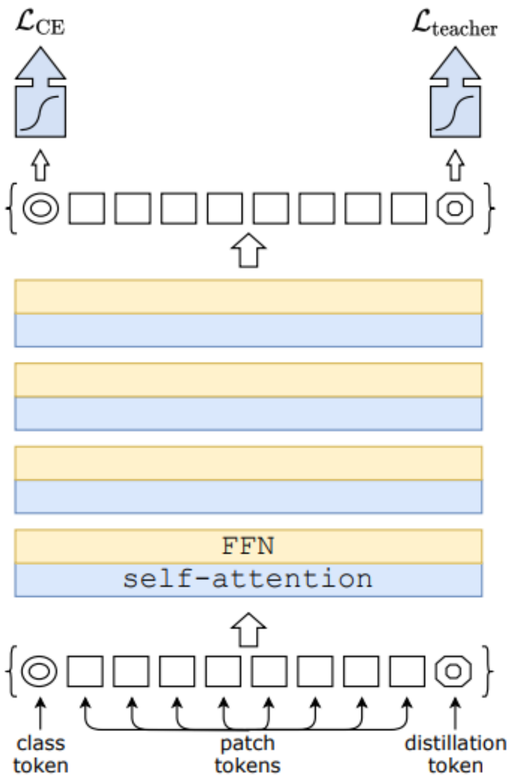


Figure 2: Our distillation procedure: we simply include a new *distillation token*. It interacts with the class and patch tokens through the self-attention layers. This distillation token is employed in a similar fashion as the class token, except that on output of the network its objective is to reproduce the (hard) label predicted by the teacher, instead of true label. Both the class and distillation tokens input to the transformers are learned by back-propagation.