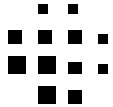


ViT

20201362
조인영

AI &
HEALTHCARE LAB

Introduction

Inductive Bias

***Inductive Bias**

Inductive Bias는 training에서 보지 못한 데이터에 대해서도 적절한 귀납적 추론이 가능하도록 하기 위해 모델이 가지고 있는 가정들의 집합을 의미

***DNN의 기본적인 요소들의 inductive bias**

- Fully connected : 입력 및 출력 element가 모두 연결되어 있으므로 구조적으로 특별한 relational inductive bias를 가정하지 않음
- Convolutional : CNN은 작은 크기의 kernel로 이미지를 지역적으로 보며, 동일한 kernel로 이미지 전체를 본다는 점에서 locality와 transitional invariance 특성을 가짐
- Recurrent : RNN은 입력한 데이터들이 시간적 특성을 가지고 있다고 가정하므로 sequentiality와 temporal invariance특성을 가짐

→ Transformer는 CNN및 RNN보다 상대적으로 Inductive bias가 낮음

Introduction

- NLP에서 사용되는 standard transformer를 이미지에 그대로 적용해 이미지 분류에 좋은 성능을 도출한 Vision Transformer(ViT)를 제안함

- ViT는 이미지를 patch로 분할한 후, 이를 NLP의 단어로 취급해 각 patch의 linear embedding을 순서대로 Transformer의 Input으로 넣어 이미지를 분류함

- ViT를 Imagenet-1k로 학습시켰을 때, 비슷한 크기의 ResNet보다 낮은 성능을 보임
  → ViT가 CNN보다 inductive bias가 낮은 것을 알 수 있음

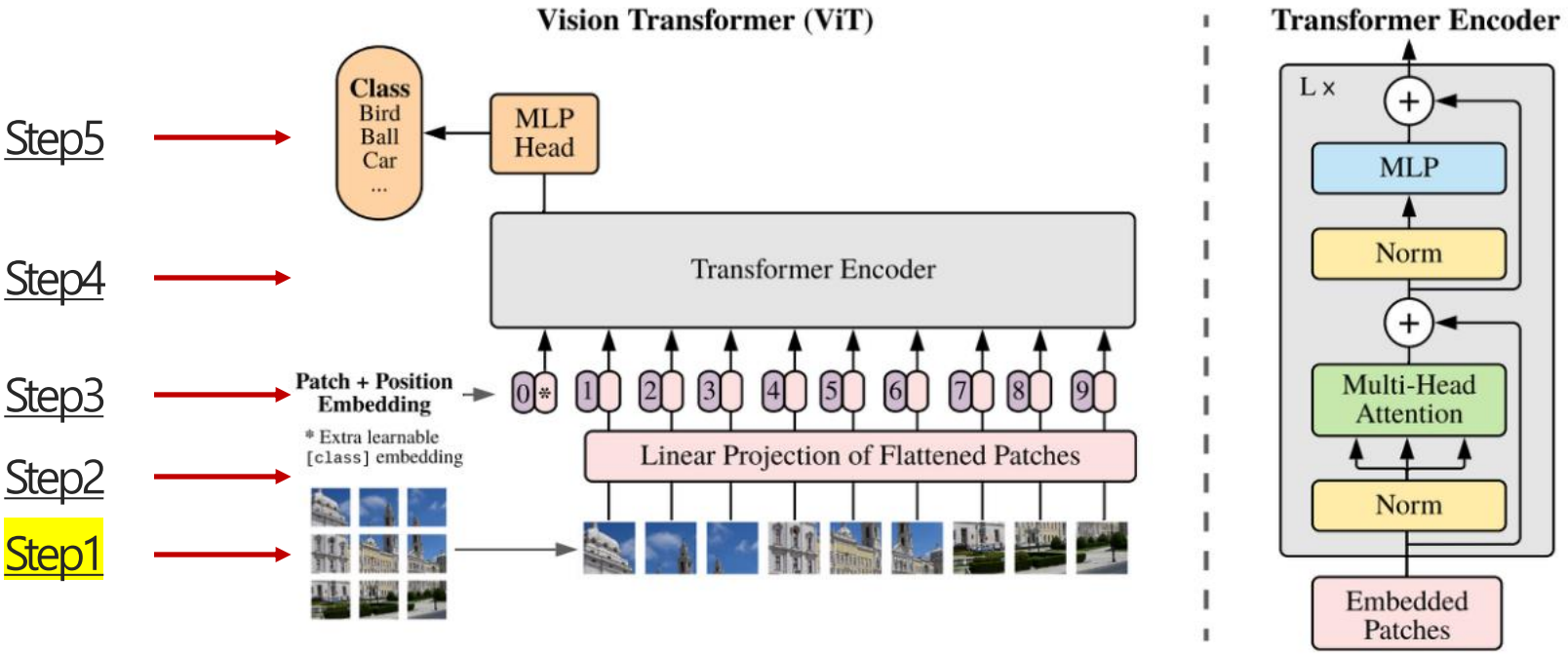
- 반면, 대규모 사이즈의 데이터인 ImageNet-21K와 JFT-300M을 통해 pre-training한 후 transfer learning 했을 때, ViT가 SOTA성능을 도출함
  → 대규모 사이즈의 학습이 낮은 inductive bias로 인한 성능 저하를 해소시킴

Method

Vision Transformer

- **ViT 모델 구조**

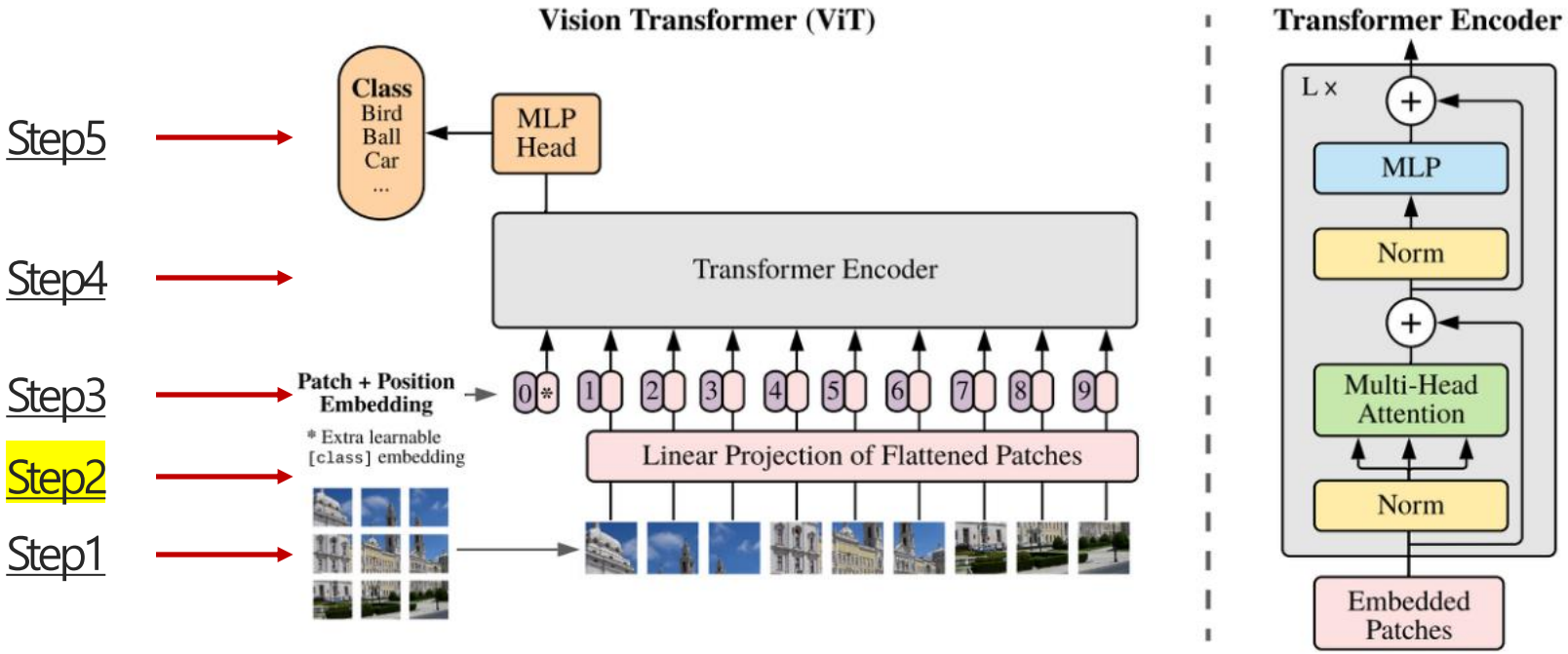
✓ Step1 – 이미지 $x \in \mathbb{R}^{H \times W \times C}$ 가 있을 때, 이미지를 (P x P)크기의 패치 N(=H x W / P^2개로 분할하여 패치 sequence $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ 를 구축함



Method

Vision Transformer

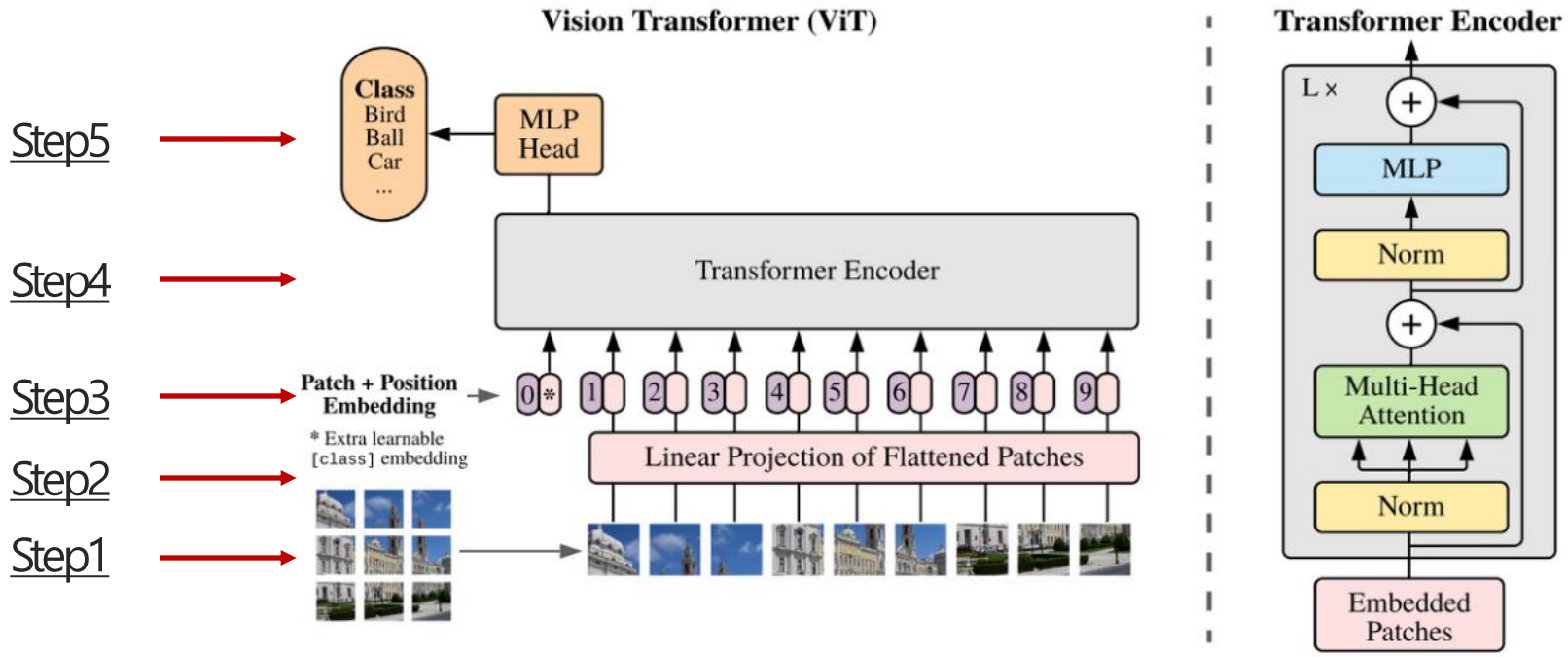
✓ Step2 – Trainable linear projection을 통해 X_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용함(step 2를 거친 후 하나의 patch가 벡터로 변환)



Method

Vision Transformer

- ✓ Step3 - Learnable class 임베딩과 패치 임베딩에 learnable position 임베딩을 더함
- ✓ Step4 – 임베딩을 Transformer encoder에 input으로 넣어 마지막 layer에서 class embedding에 대한 output인 image representation을 도출함
- ✓ Step5 – MLP에 image representation을 input으로 넣어 이미지의 class를 분류함

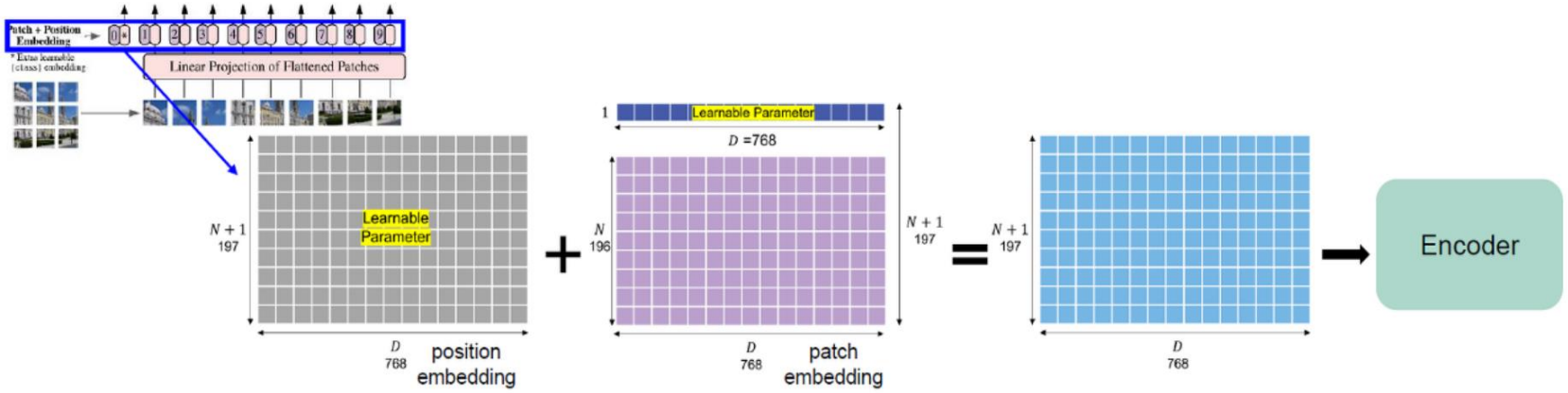


Method

Vision Transformer

- **Positional Embedding**

- ViT에서는 4가지 position 임베딩을 시도한 후, 가장 효과가 좋은 1D Position 임베딩을 ViT에 사용함
- 가장 첫번째 패치 임베딩 앞에 학습 가능한 임베딩 벡터를 붙여줌
  (이는 추후 이미지 전체에 대한 표현을 나타내게 됨)
- N+1개의 학습 가능한 1D 포지션 임베딩 벡터를 만들어준 후, 이를 각 이미지 패치 벡터와 합해줌
- 만들어진 임베딩 벡터를 Encoder에 입력

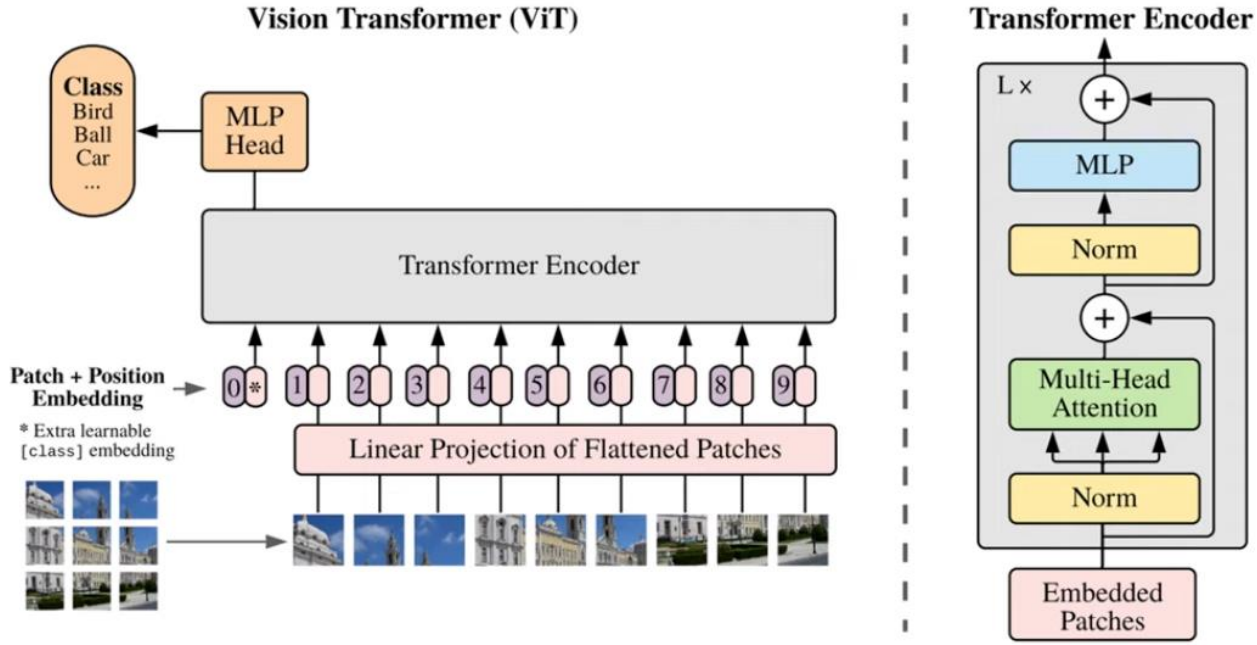


Method

Vision Transformer

- **Transformer Encoder**

- ViT는 Multi-head Self Attention과 MLP block으로 구성되어 있음
- MLP는 2개의 layer를 가지며, GELU activation function을 사용함
- 각 block의 앞에는 Layer Norm을 적용하고, 각 block뒤에는 residual connection을 적용함



Transformer

Transformer Architecture

Transformer

Transformer Architecture

Transformer

Transformer Architecture