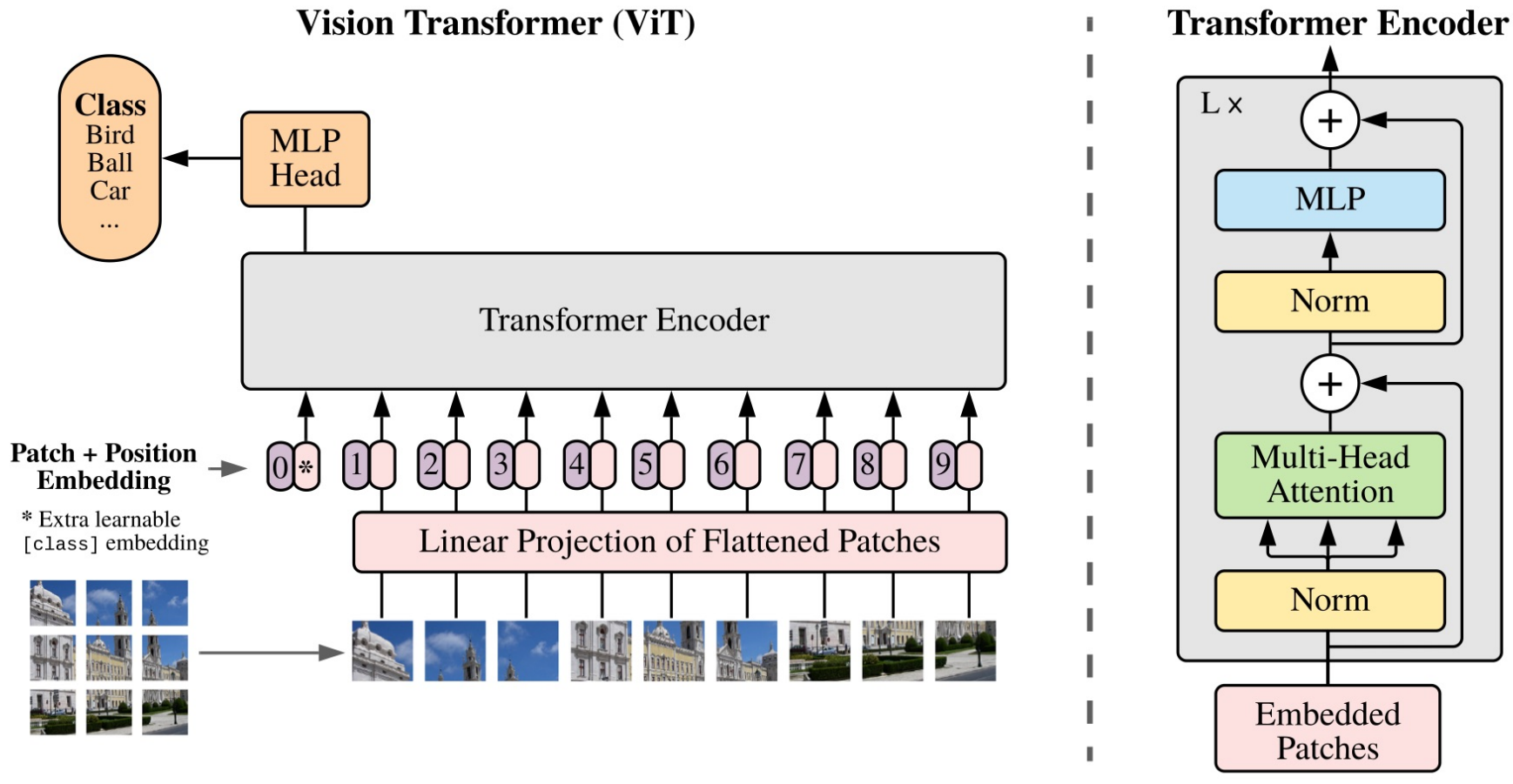
Vision Adv Assignment

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Vision Transformer

- 이미지 분류를 위함



[Embedded Patches]

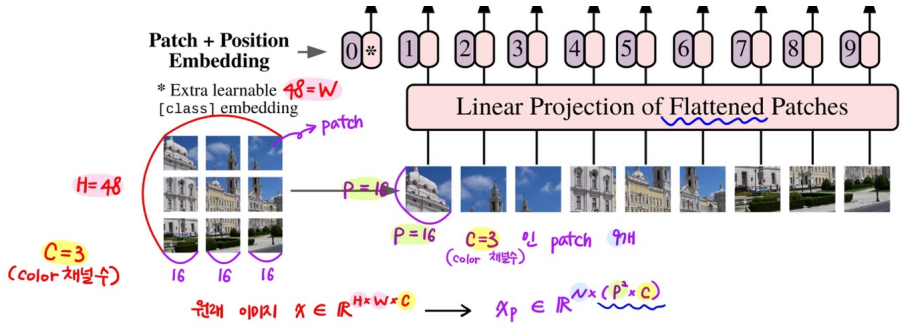
1. Input 이미지를 9개의 **patch**로 나눠준다.
2. 각각의 patch들은 선형변환을 통해 **embedding** 된다.

✔️ linear projection of Flattened Patches

- patch 크기가 P 일때, 하나의 이미지 x ∈ R^{HxWxC} 는 각 patch가 1차원 텐서로 펼쳐져서(flatten) xp ∈ R^{Nx(p^2xC)} 시퀀스가 된다.

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

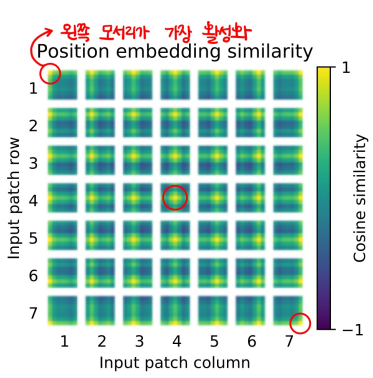
  - Ex. 이미지 크기가 H=W=48이고, patch 크기가 P=16 일때 3x3=9개의 16x16x3=768 차원의 텐서로 이루어진 시퀀스가 되는 것임.
- 이후에는 시퀀스의 각 요소 별로 임베딩을 위한 선형변환을 수행



✔️ 이미지 임베딩(Image Embedding)

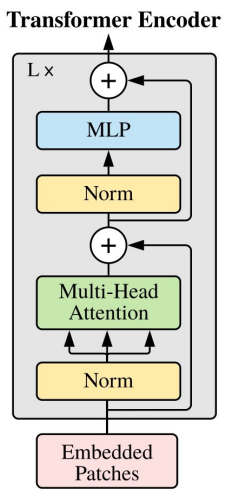
- 이미지를 저차원 공간에 수학적으로 표현한 것
- 이미지의 특징과 중요한 feature를 잡아낸다.
- patch 임베딩 결과 9 x 768 크기의 패치 임베딩 행렬 도출됨.

3. 각각의 patch는 위치가 중요하므로 가장 앞부분에 position embedding을 해준다.



4. embedding된 결과는 **transformer encoder의 입력 시퀀스**가 된다.

[Transformer Encoder]



Norm

5. embedding 된 patch는 normalization 시킨다.

Multi-Head Attention

6. Self Attention 구조에 들어가기 위해 해당 vector에 가중치를 곱하여 query, key, value로 나눠준다.

⊕

7. Skip connection 을 통해 MHA를 통과한 output과 통과하지 않은 원래 값을 더해 기존의 값을 보존해준다.

Norm

8. 그 결과를 normalization layer에 통과시킨다.

MLP(Multi Layer Perceptron)

9. multi layer perceptron에 통과시킨다.

⊕

10. MLP를 통과한 값과 MHA를 통과한 기존의 값을 더해준 Skip Connection을 통해 최종 output을 출력한다.

[MLP Head]

11. Transformer Encoder의 출력값은 MLP를 통해 어떤 이미지인지 분류된다.

참고자료

2010.11929.pdf