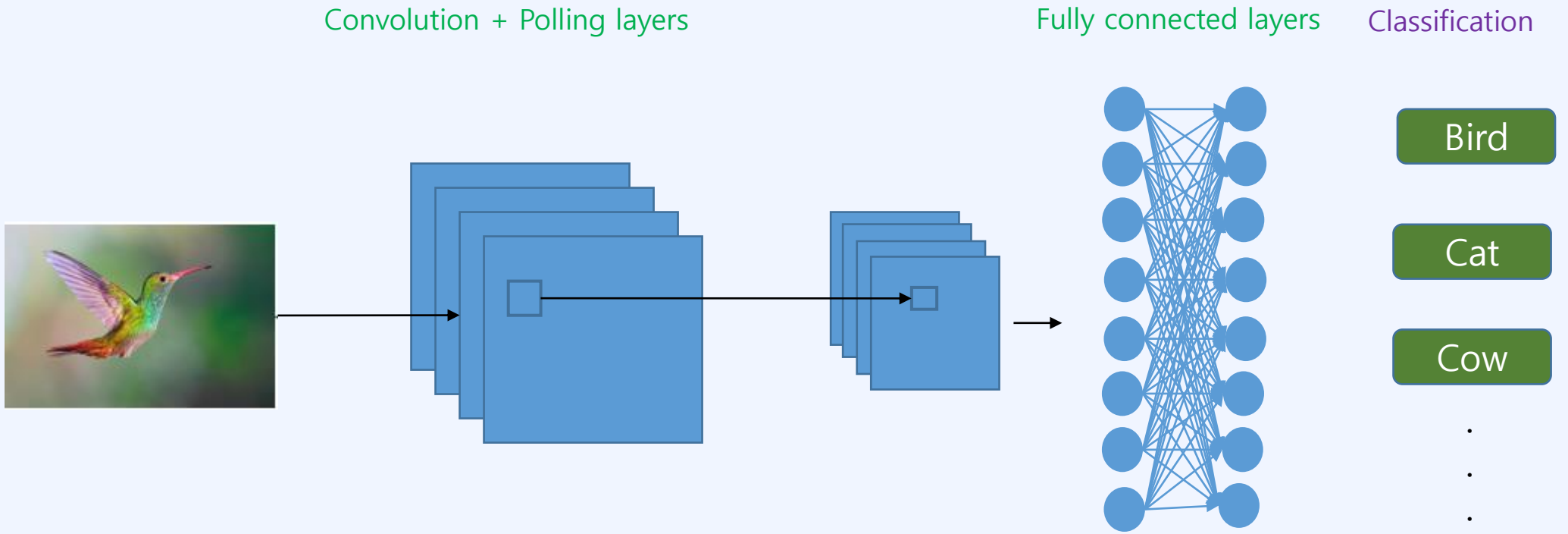
Part 4
SOTA model

Ch1. ViT

An image is worth 16x16 words - Transformers for image recognition at scale

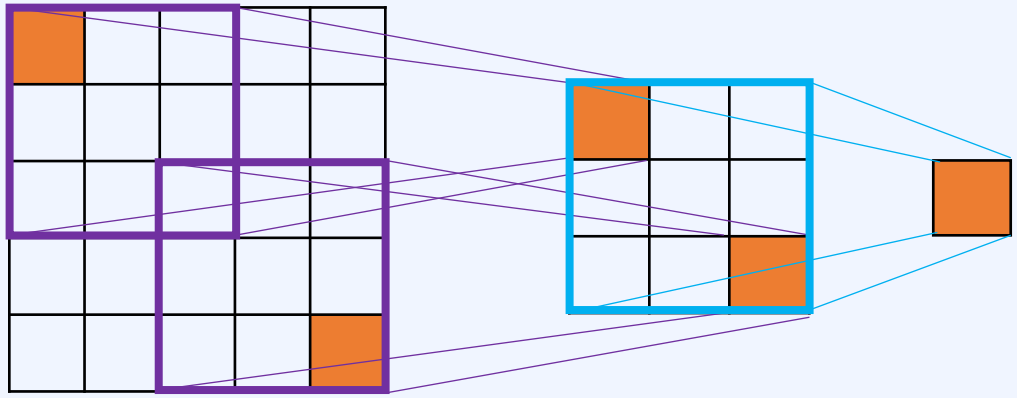
CNN to ViT

- CNN(Convolutional Neural Network)
  - Computer vision분야에서 많이 사용
  - Input image의 공간정보를 유지한 채 학습

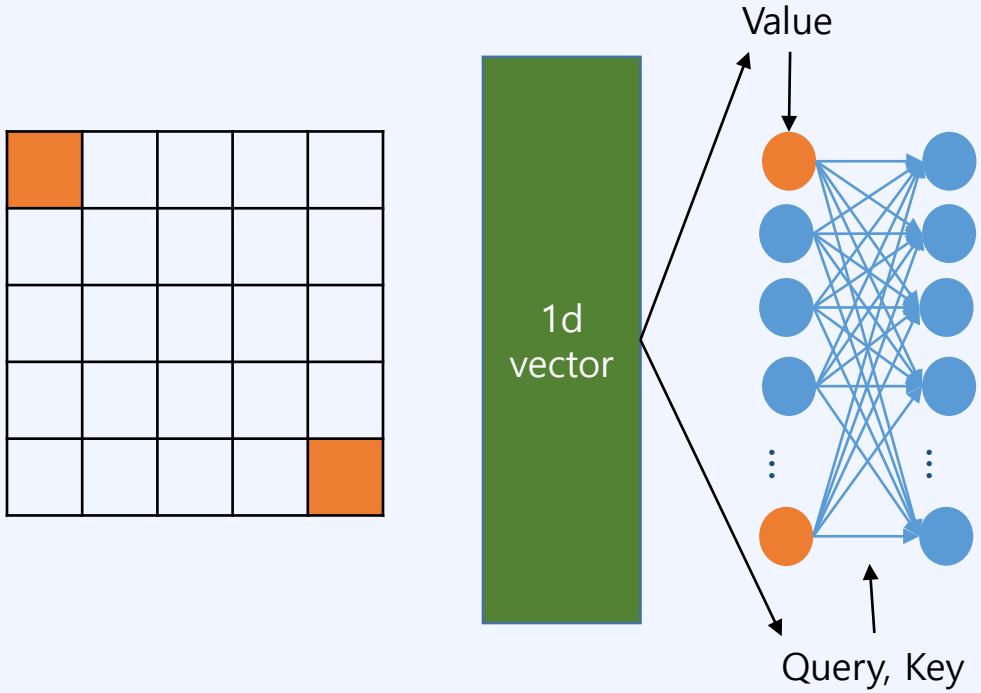


CNN to ViT

- Transformer & CNN
  - CNN – Image 전체의 정보를 압축하기 위해 여러 개의 layer를 통과
  - Transformer – 하나의 layer로 전체 image 정보를 압축



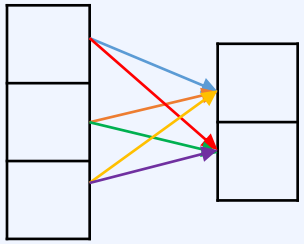
CNN



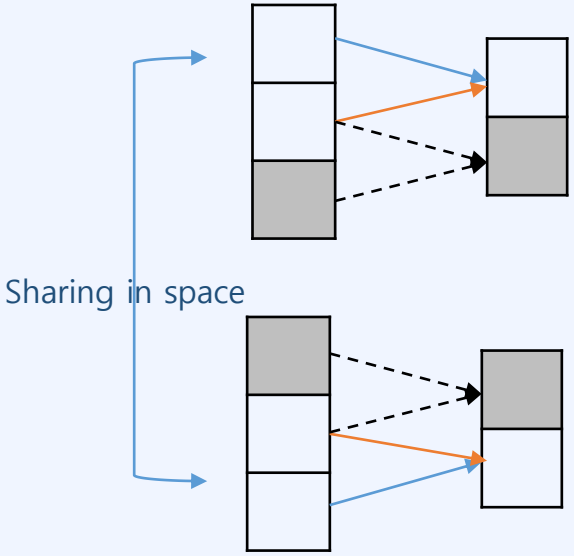
Transformer

Transformer

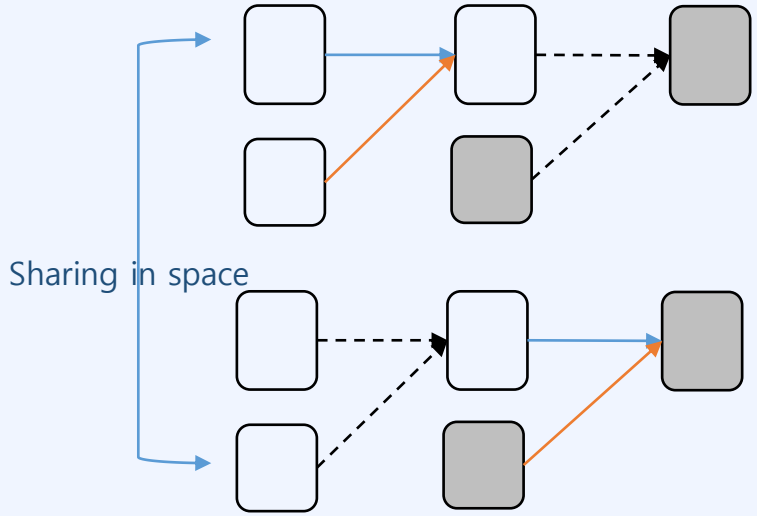
- Inductive bias
  - 주어지지 않은 입력의 출력을 예측



Fully connected



Convolution



Recurrent

Transformer

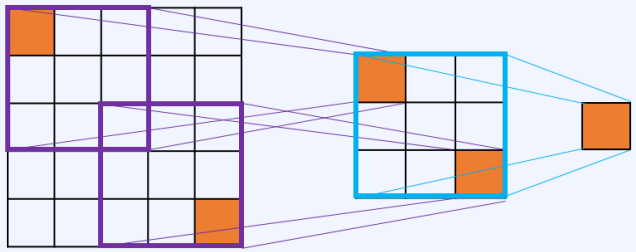
- Inductive bias

CNN

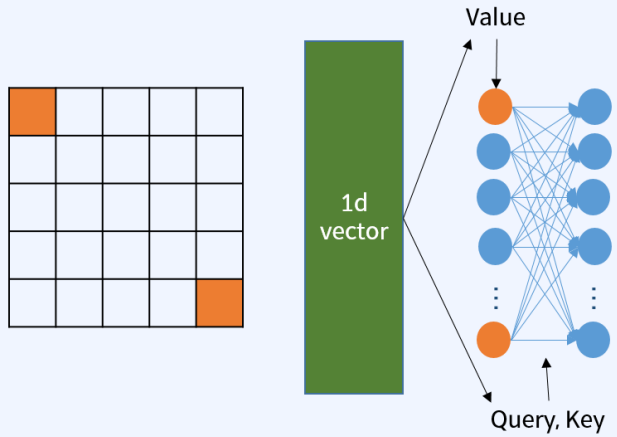
- Convolution filter 사용
- 지역적인 정보 유지o
- 학습 후, 고정된 Weight을 사용

Transformer

- 임베딩에 의한 벡터 변환 후, Self attention
- 지역적인 정보 유지x
- 학습 후에도 input vector에 따라 Weight이 달라짐
- Inductive bias ↓



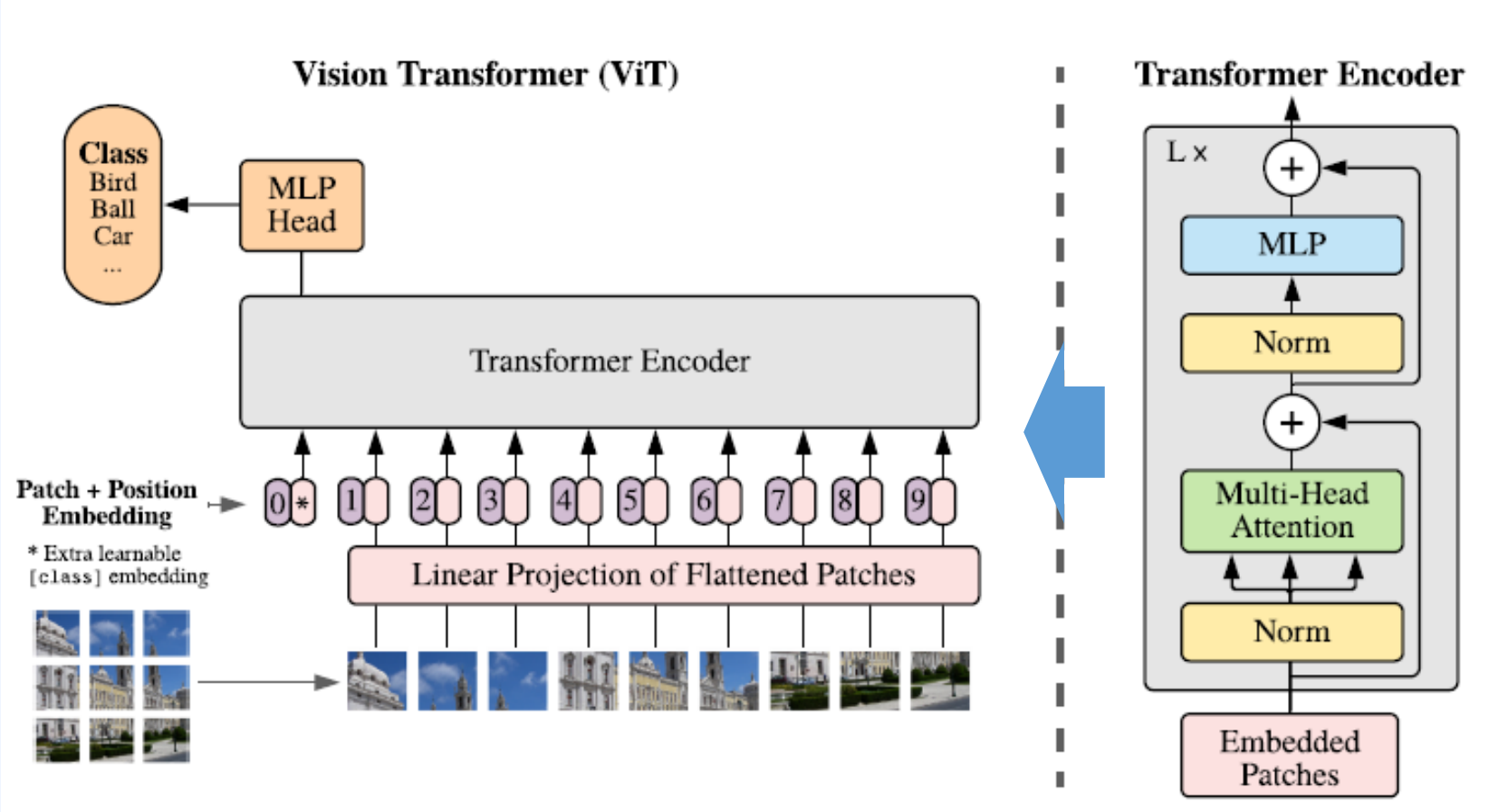
CNN



Transformer

ViT(Vision in Transformer)

- Architecture

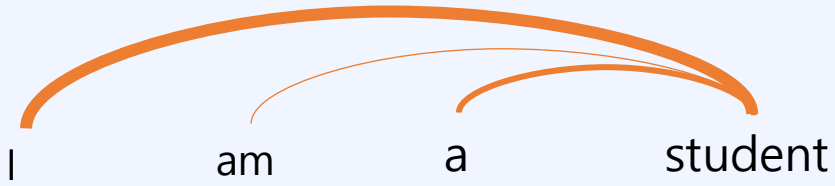


ViT 동작

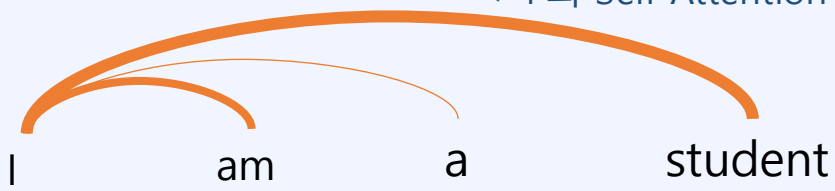
Self attention

- NLP

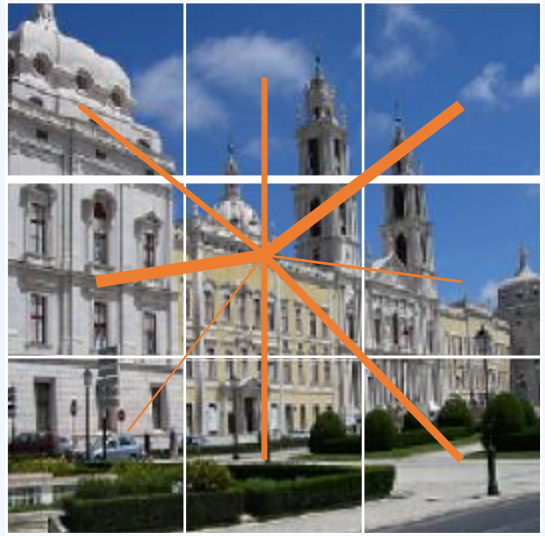
< 'student'의 Self-Attention >



< 'I'의 Self-Attention >

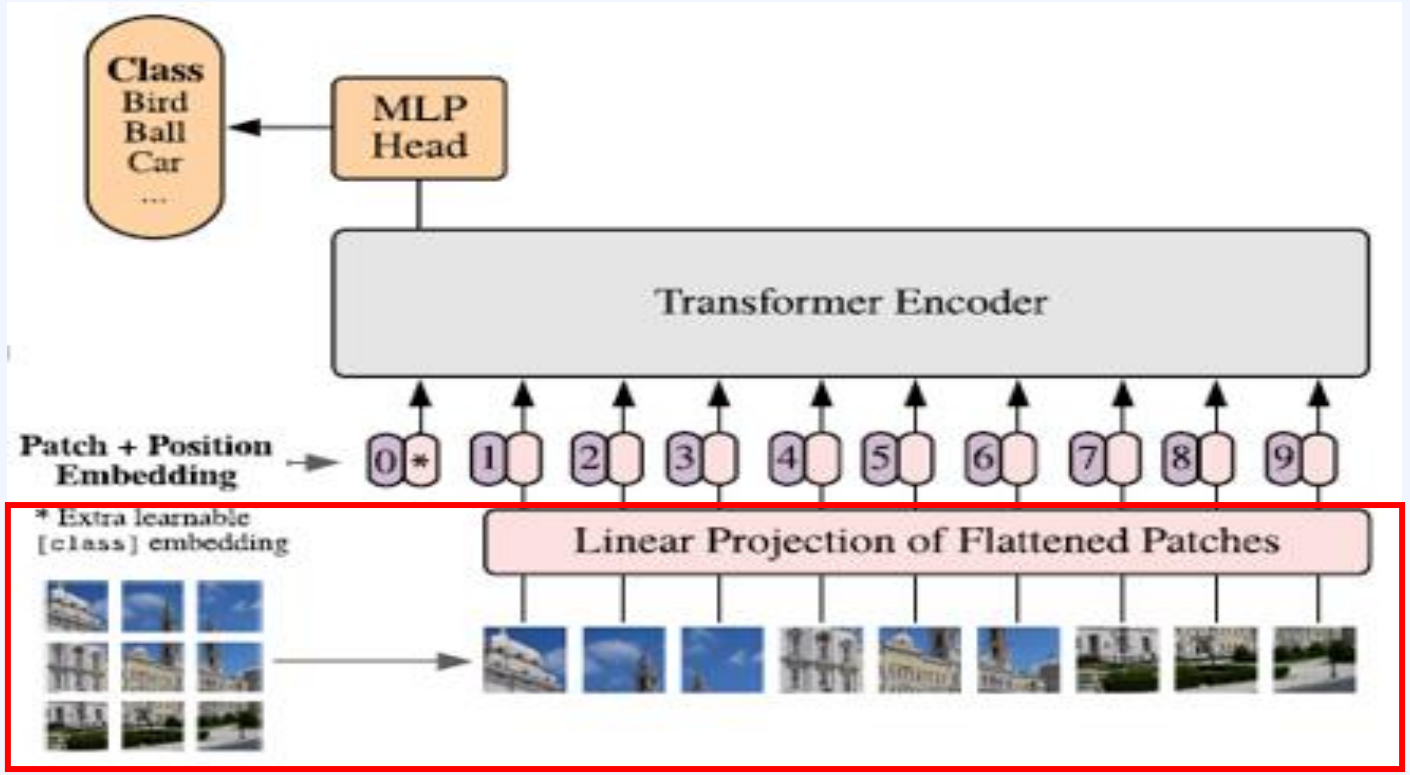


- Vision



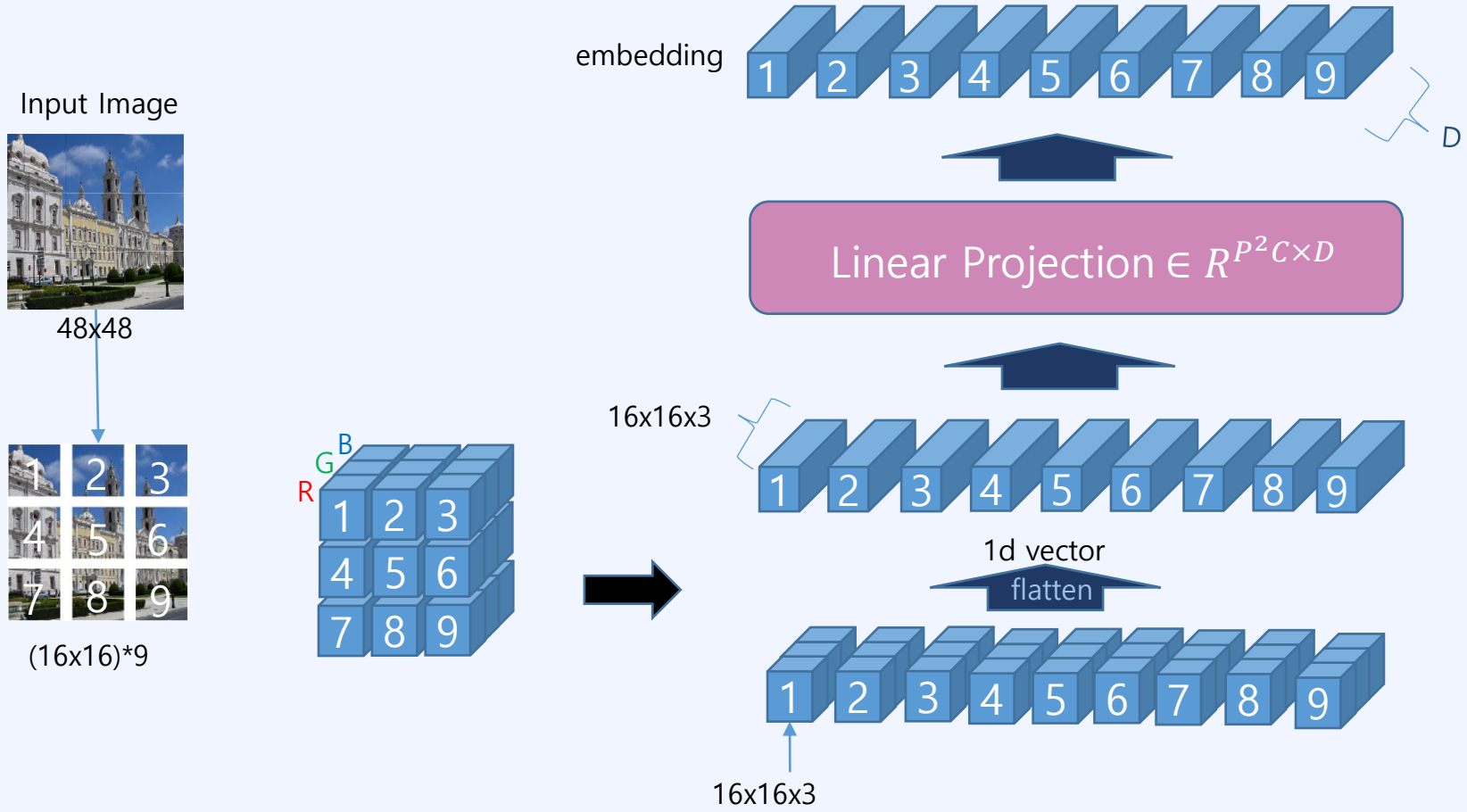
ViT 동작

- Step 1 - Patch embedding



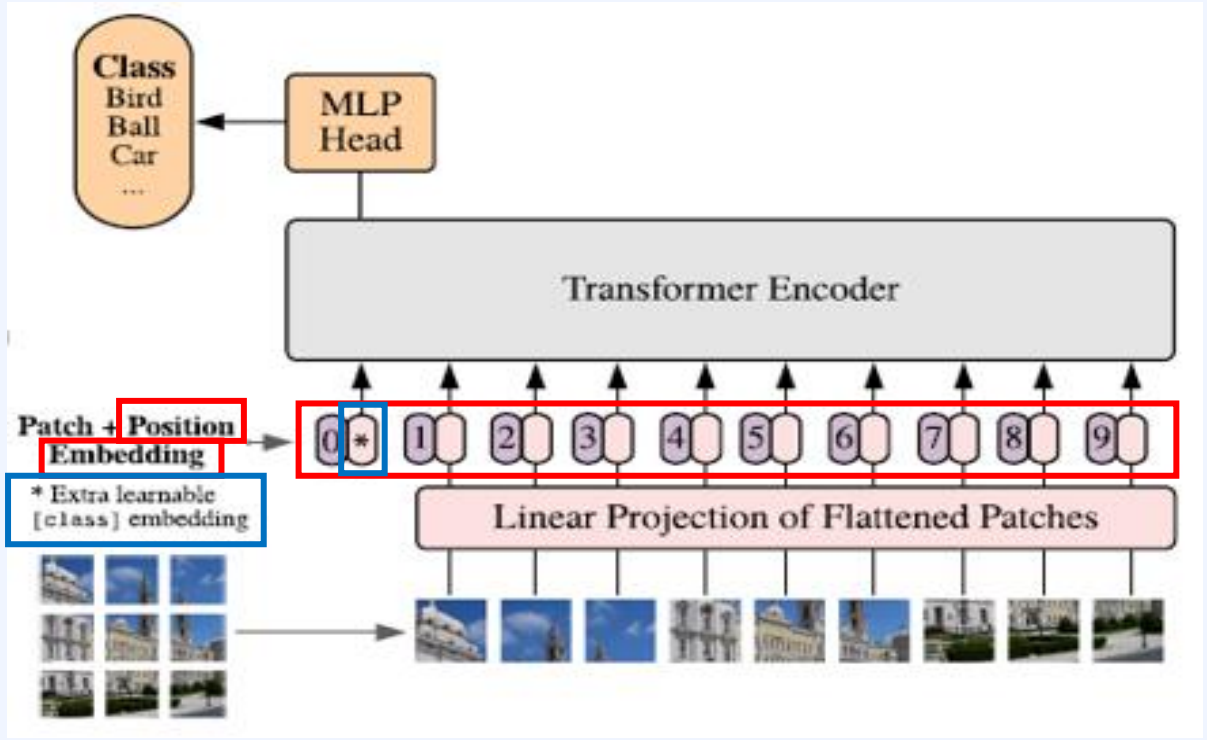
ViT 동작

- Patch embedding(example-ViT/16)



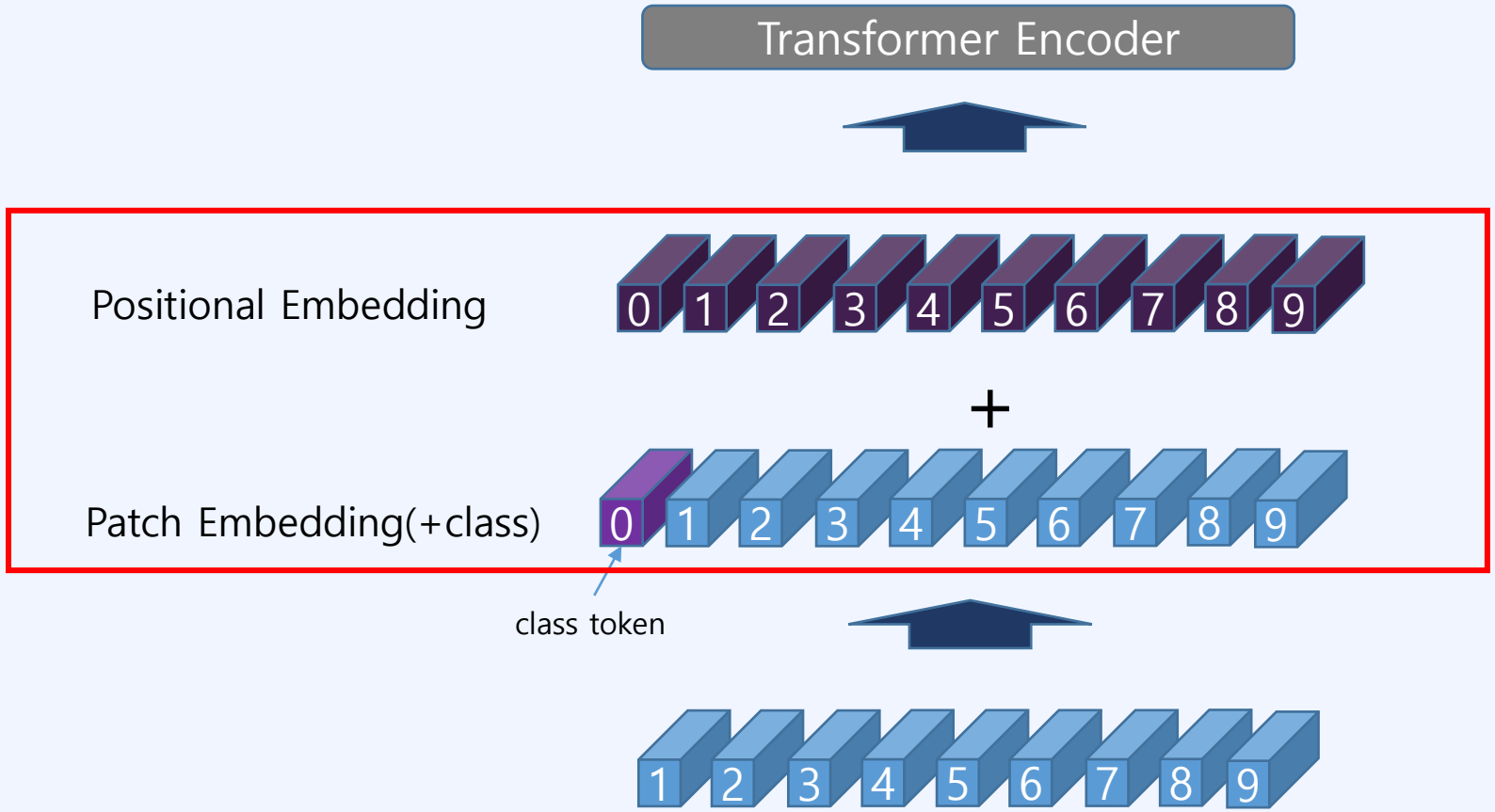
ViT 동작

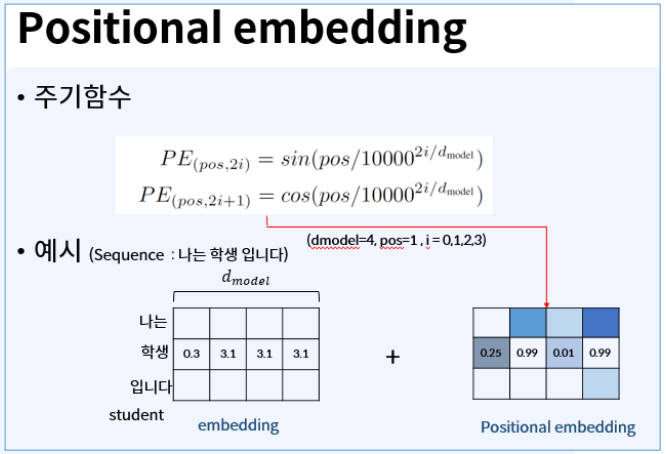
- Step2 - Embedding patch + Positional Embedding



ViT 동작

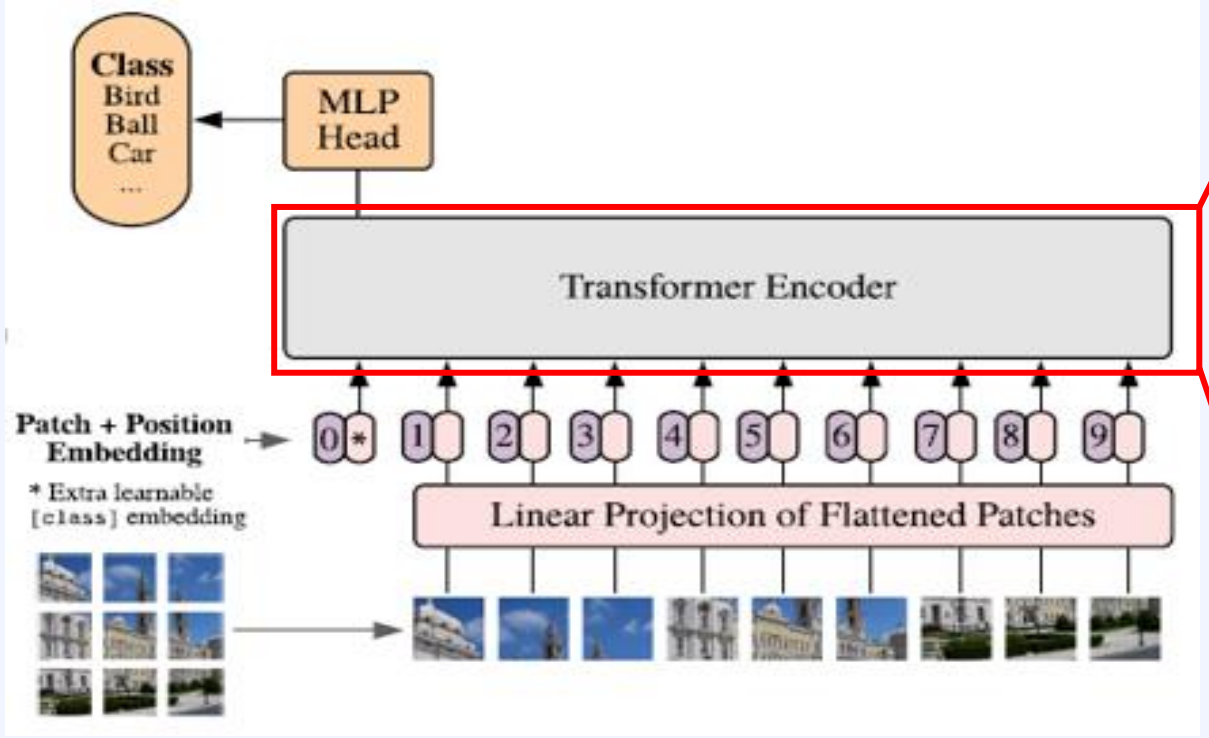
- Patch embedding + Positional Embedding (example-ViT/16)

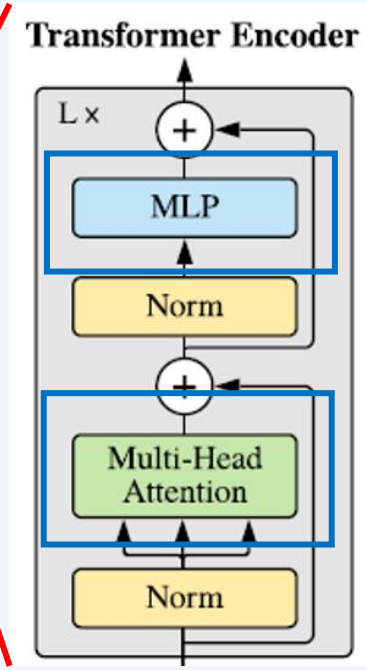




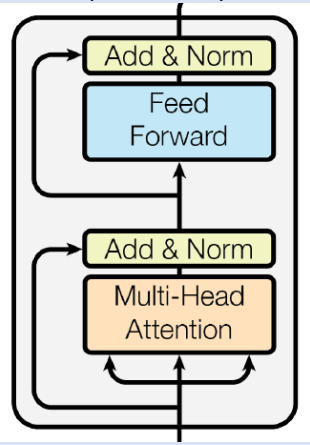
ViT 동작

- Step 3 – Transformer encoder





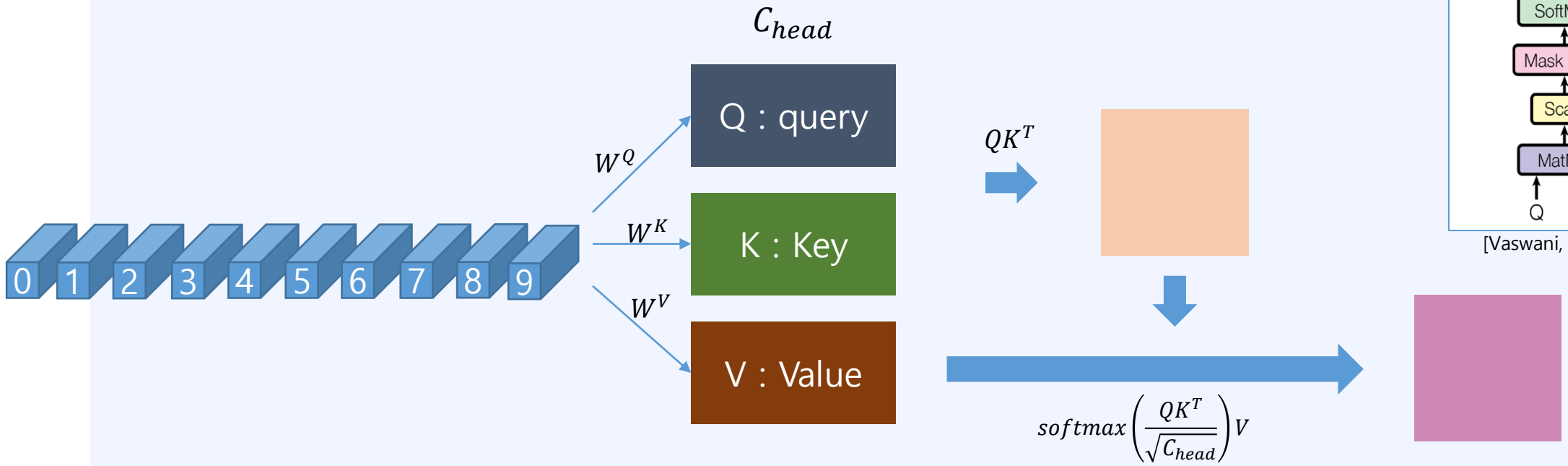
Transformer encoder (standard)

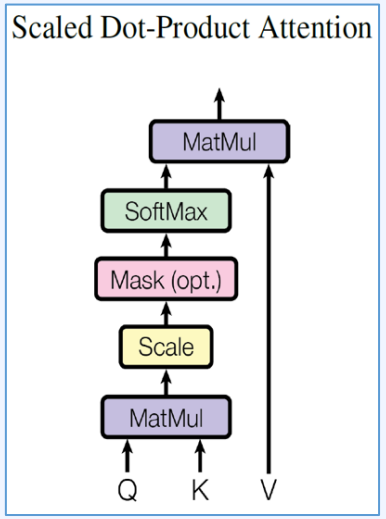


[Vaswani, A.(NIPS'17)]

ViT 동작

- Multi-Head Attention
  - Scaled Dot-Product Attention

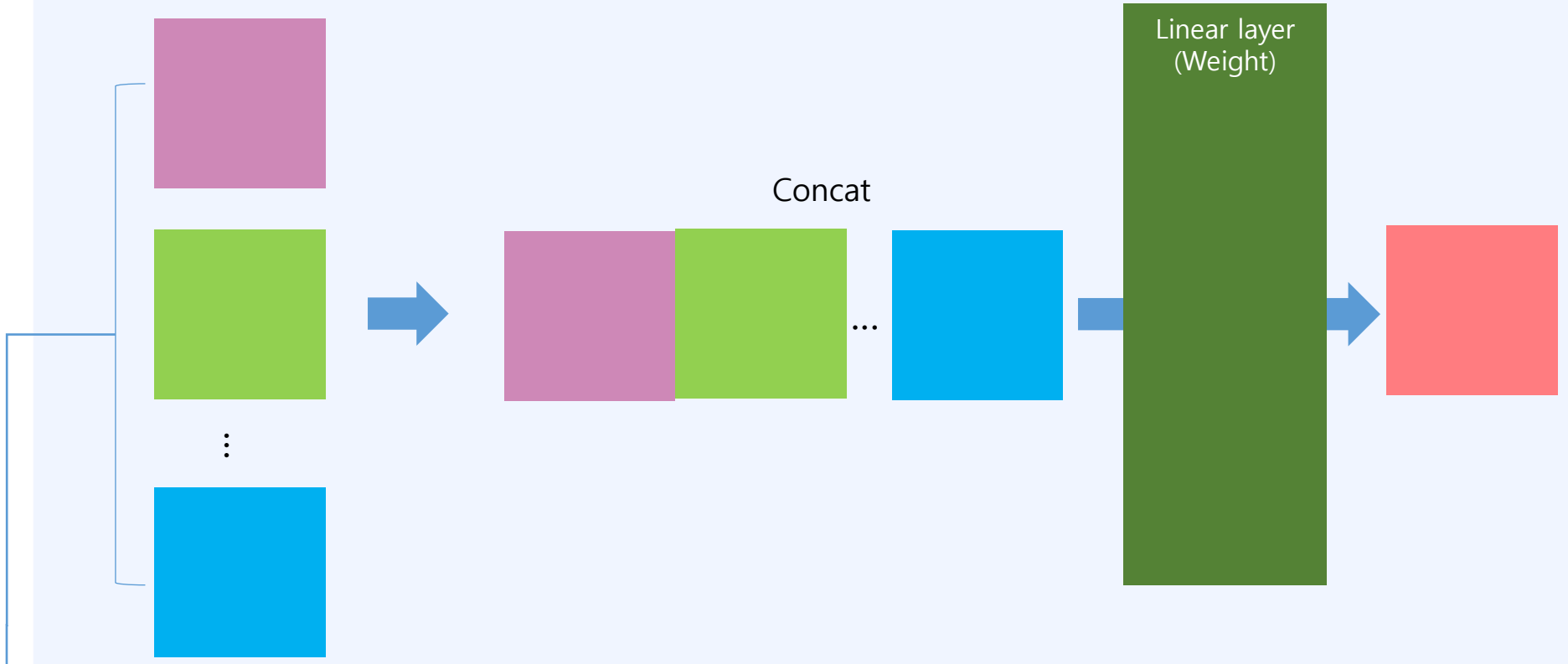


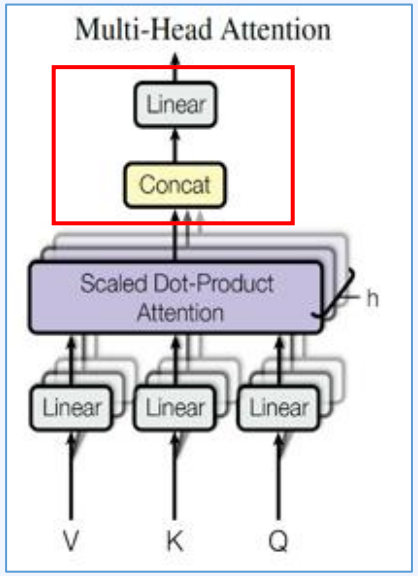


[Vaswani, A.(NIPS'17)]

ViT 동작

- Multi-Head Attention

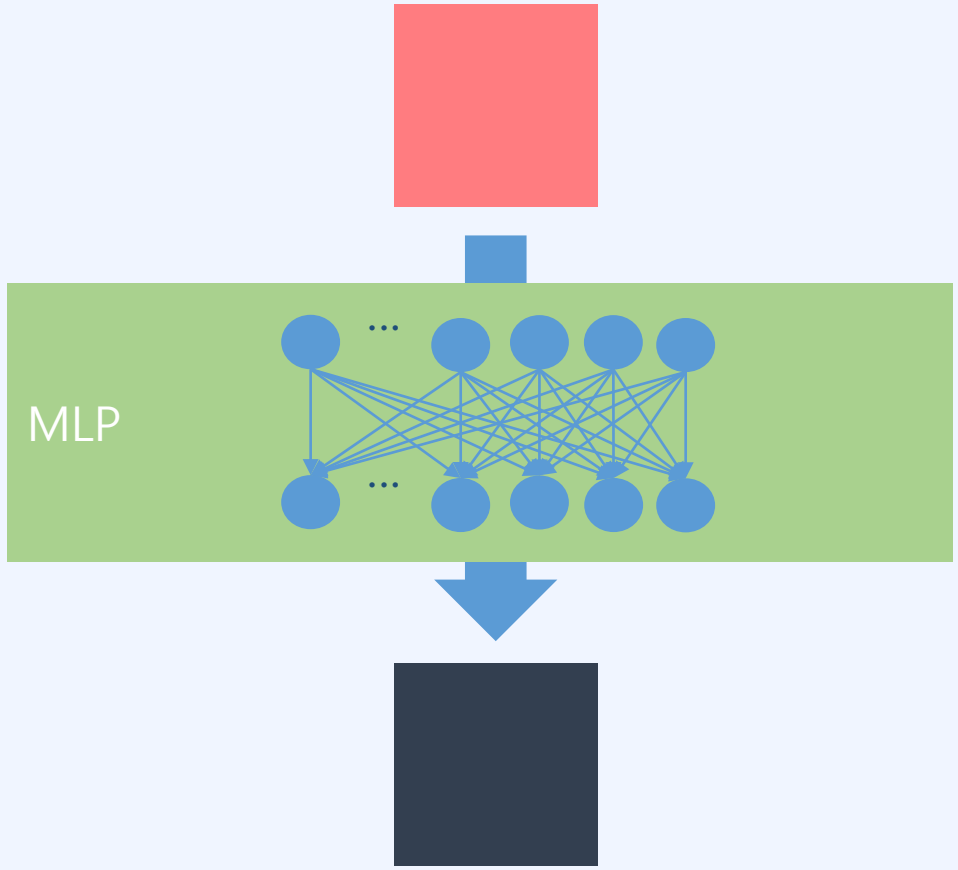


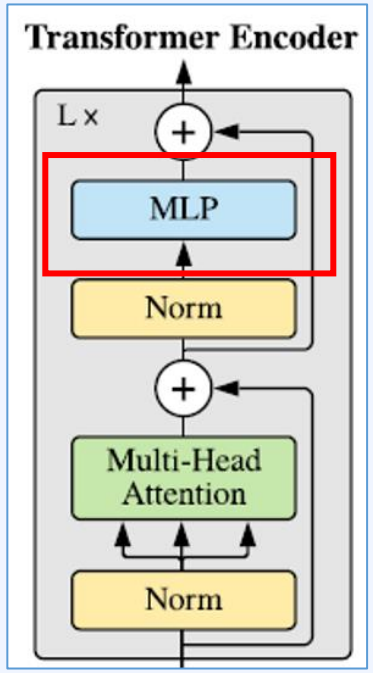


[Vaswani, A.(NIPS'17)]

ViT 동작

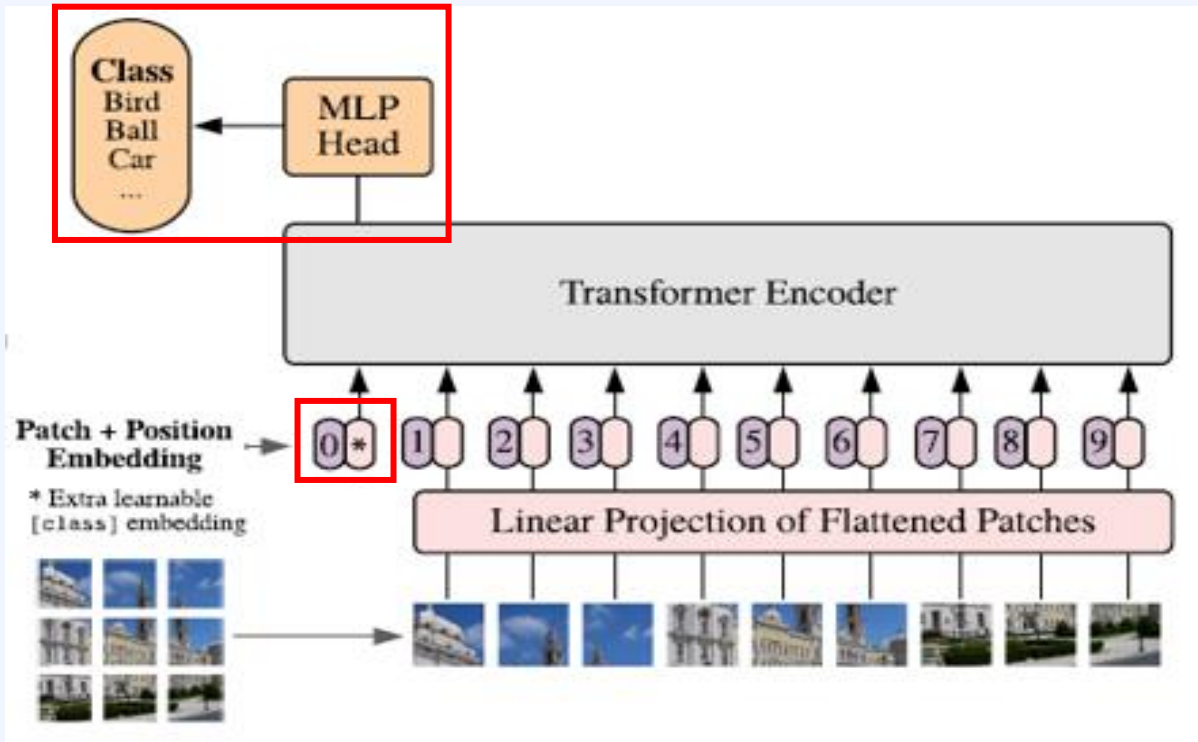
- MLP





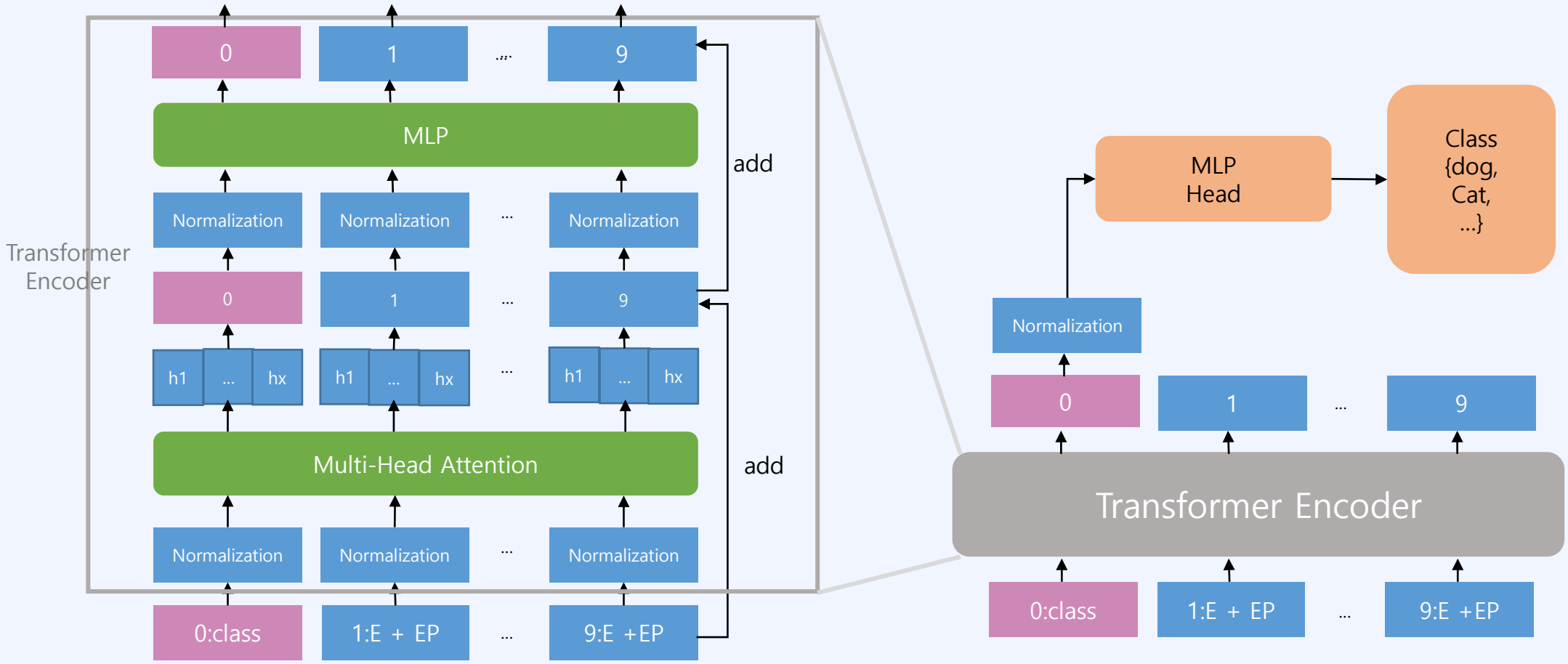
ViT 동작

- Step 4 – MLP Head



ViT 동작

- MLP Head



Summary