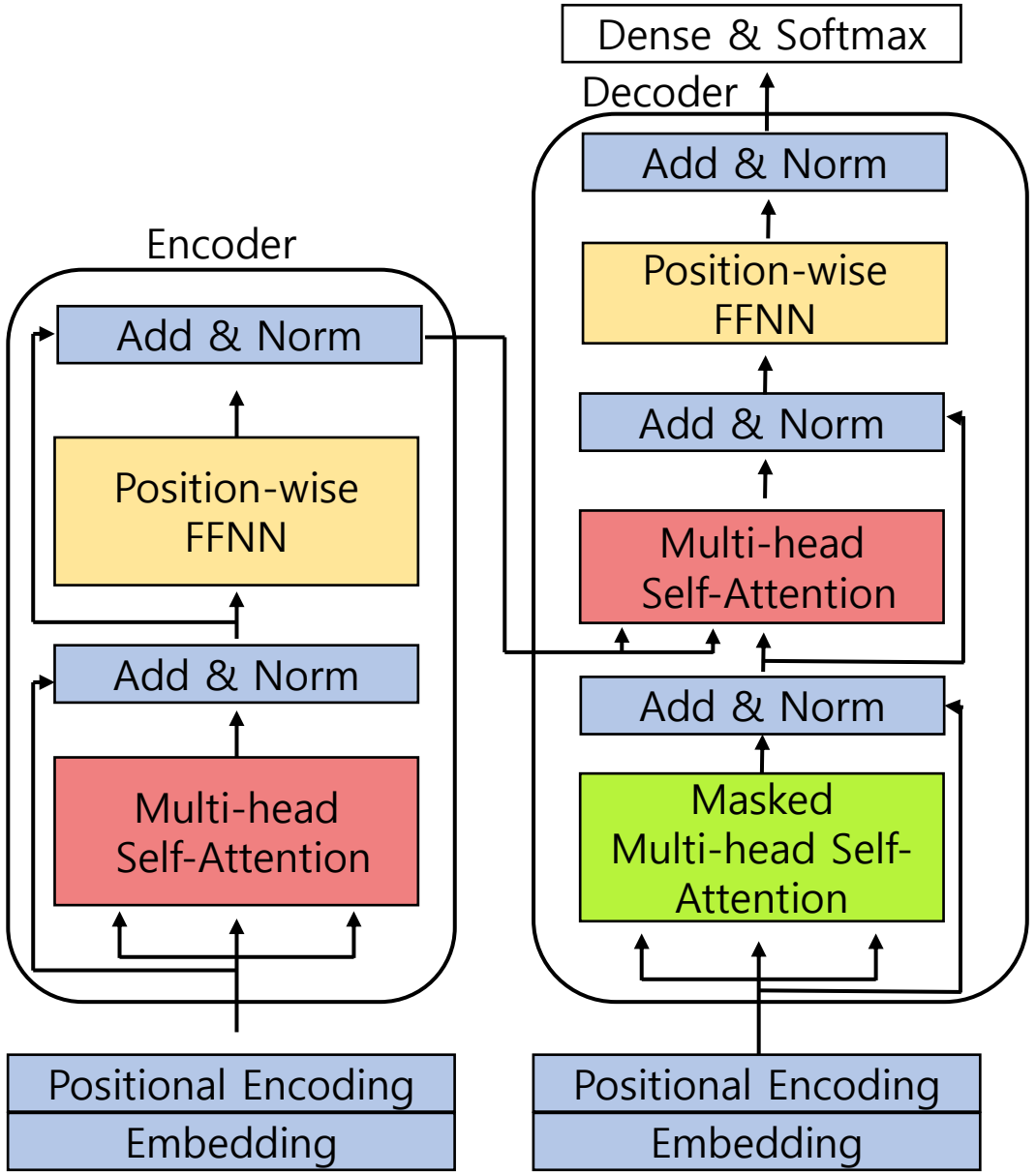
Vision Transformer - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Written in 2022.08.27

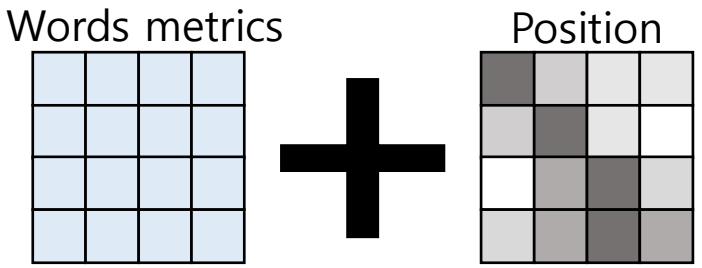
Research Background

- Vision transformer는 transformer 모델을 vision task에 적용시킨 것임
- Transformer는 attention과 Feedforward Neural network를 사용하여 만든 모델임



Research Background

- Transformer 모델 이전의 방법들은 입력되는 문장을 순서대로 입력되지만, Transformer에서는 Positional Encoding을 통해 위치 정보를 추가함
- Embedding된 단어와 위치 정보를 더하여 위치 정보를 추가함. 이때 위치 정보는 짝수 번째에는 sin, 홀수 번째에는 cos 함수를 사용함

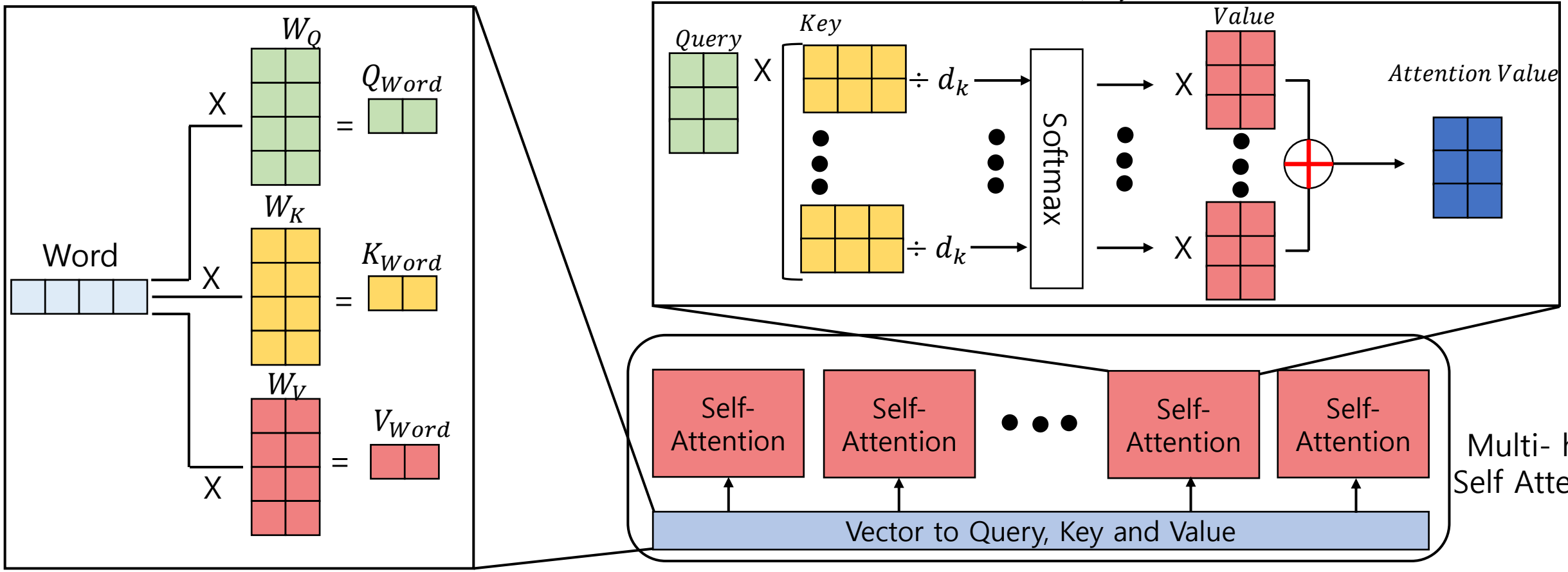


$$PE(pos,2i)=\sin(pos/10000^{2i}/d_{model})$$
$$PE(pos,2i+1)=\cos(pos/10000^{2i}/d_{model})$$

Research Background

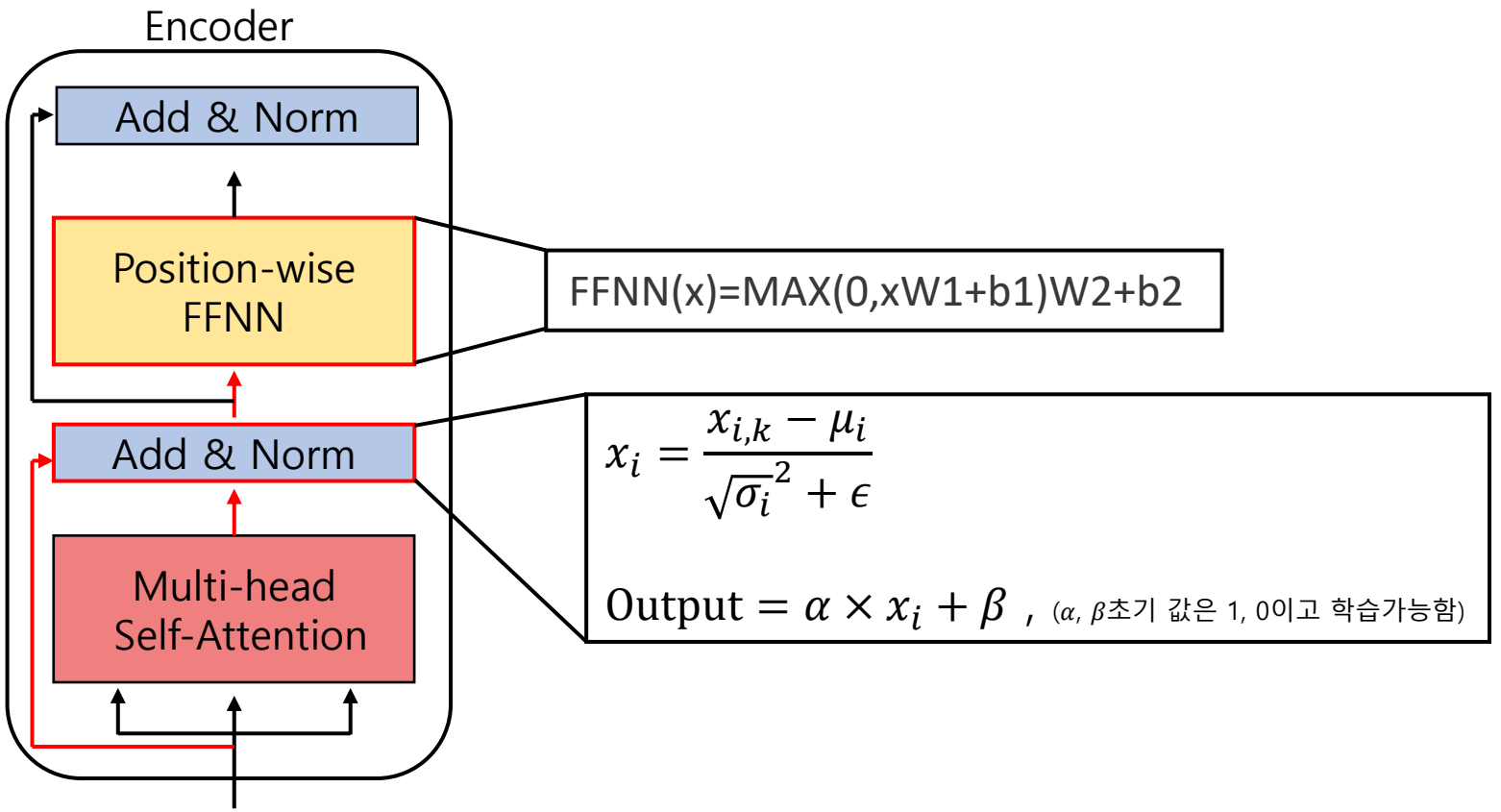
- Attention 분석이 대상이 되는 단어(Query)와 전체 단어(Key)의 유사도를 구하는 것임
- Self Attention은 Query가 Key안에 있고, 이 때의 유사도를 구함
- Multi-head Self Attention은 Self Attention을 병렬로 나눠 진행하여 각각 다른 정보를 얻음

*Value: 키의 의미를 나타내는 가중치 벡터, * d_k: key vector의 차원



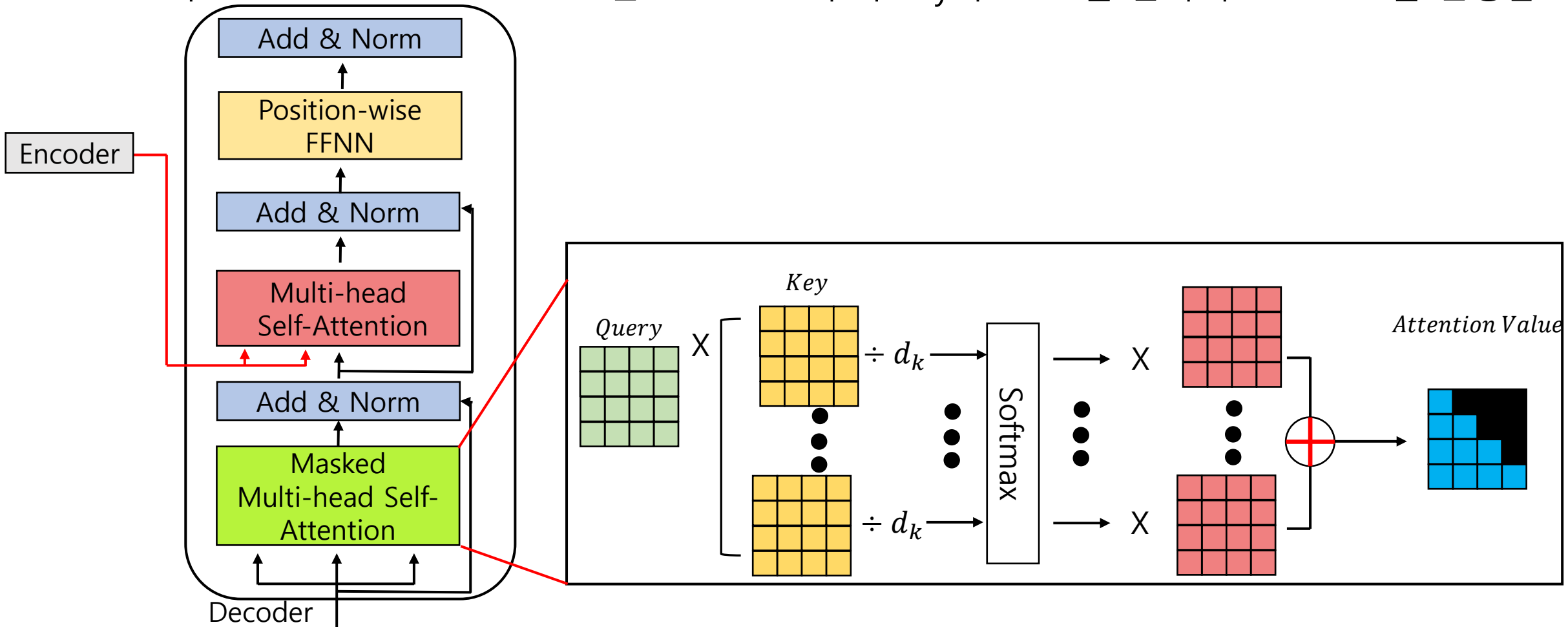
Research Background

- ResNet과 같이, Multi-head Self-Attention으로부터 나온 값은 Multi-head Self-Attention 이전의 값과 더해지고, 평균과 분산을 이용한 normalize를 함
- 그후 Position-wise FFNN(Feed Forword Neural Network)를 거침



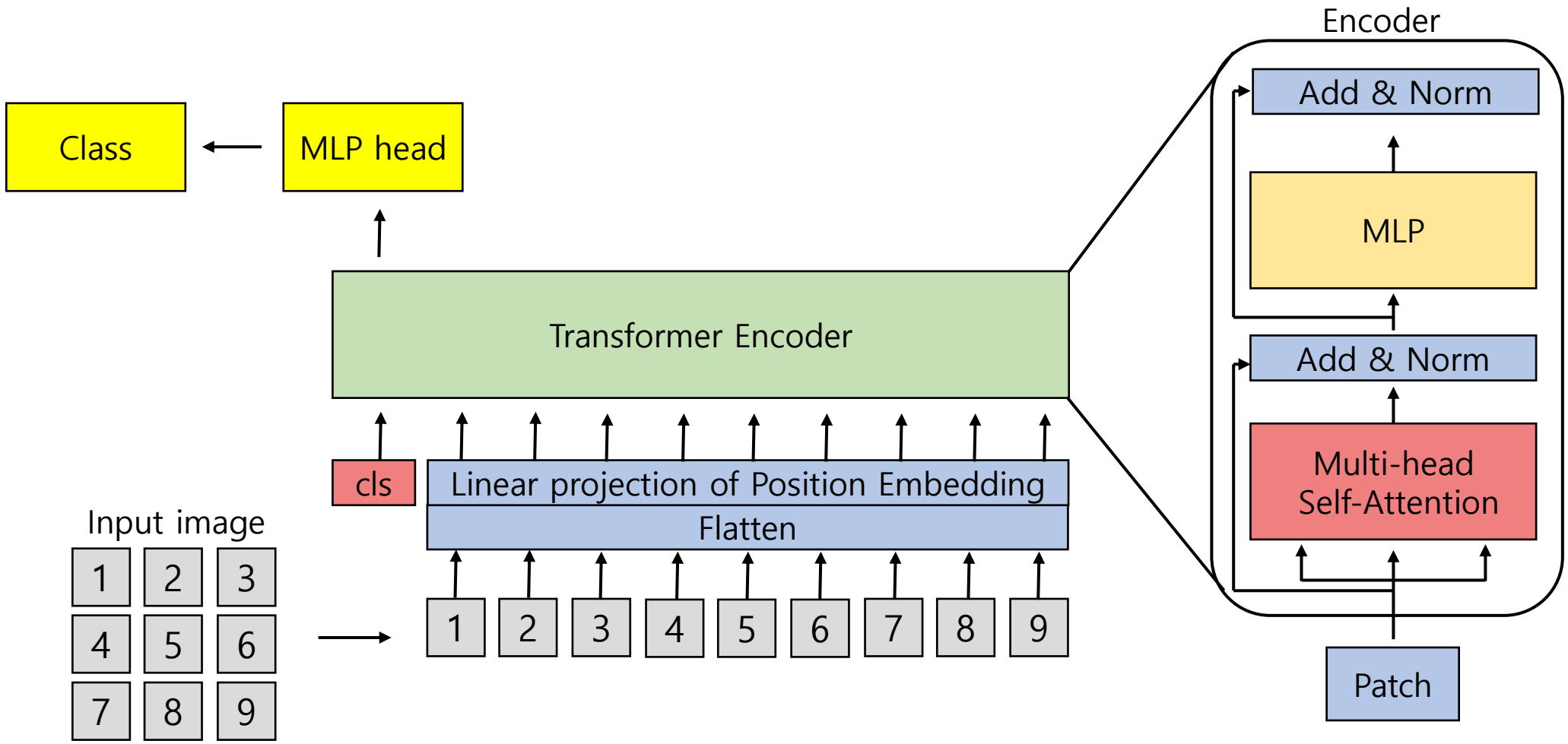
Research Background

- Transformer 모델은 미래 시점의 단어까지도 참고할 수 있는 현상이 발생하고, 이를 방지하기 위해 룩-어헤드 마스크(look-ahead mask)를 도입함.
- 이는 기존 Multi-head Self-Attention과 같지만 Attention Value 행렬의 미래의 단어에 해당하는 행을 매우 작은 값을 넣는다
- Decoder의 Multi-head Self-Attention은 Encoder로부터 Key와 Value를 받아서 Attention을 진행함



Model

- 기존 Transformer의 Encoder만을 사용하고, BERT의 class Token처럼, 학습 가능한 Embedding patch를 추가
- 입력은 하나의 이미지를 패치로 만든 다음 Flatten하고 위치 정보를 위해 Position embedding을 진행함



Result

- ImageNet과 같은 중간 크기의 데이터셋에 이 모델을 적용하였을 때는 좋지 않은 성능을 보였으나 충분히 큰 스케일에서 vision transformer를 사전 학습한 결과, 더 적은 데이터셋을 가진 하위 태스크에 전이 학습하여 좋은 성능을 얻을 수 있었음

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

<사전 학습된 데이터에 따른 benchmark 데이터에 대한 성능>

한줄평

- 많은 양의 데이터를 통해 학습해야 하긴 했지만 RNN기반의 모델이 Computer vision 분야에서 좋은 성능을 낼 수 있다는 점이 놀라웠다.