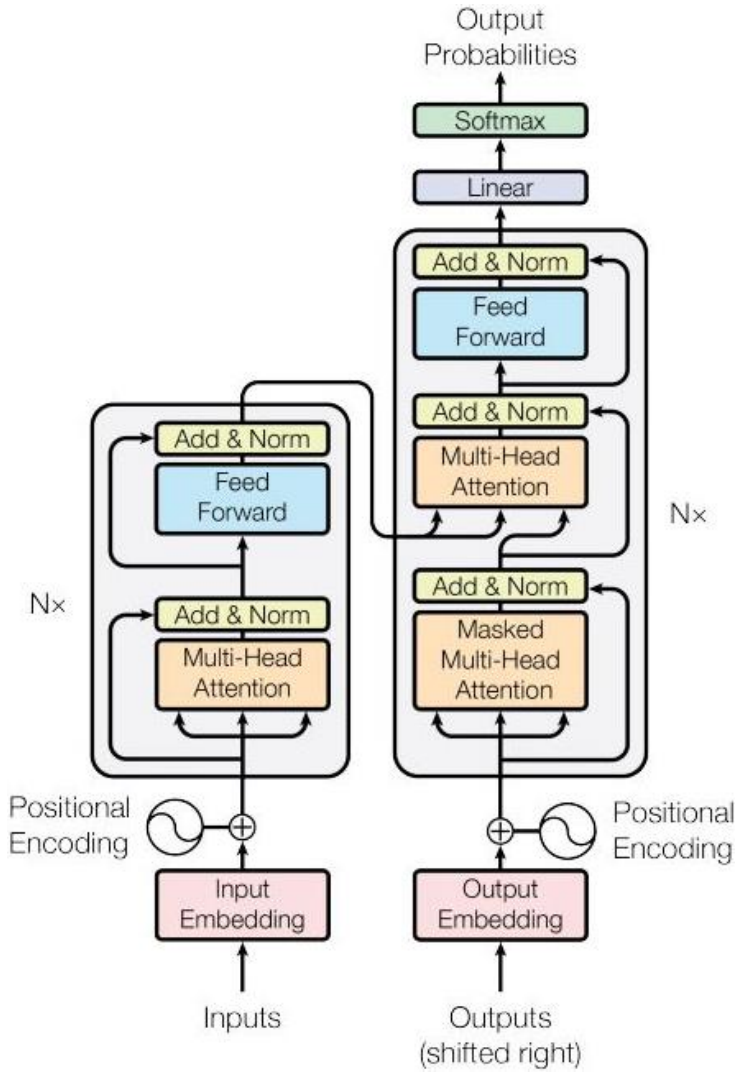
Vision Transformer - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Motivation

- 입력 시퀀스를 하나의 벡터 표현으로 압축하고, 디코더는 이 벡터 표현을 통해서 출력 시퀀스를 만들어 낸다. 이 때 **어텐션만**으로 인코더와 디코더를 만든 것이 Transformer이다.

- Transformer의 입력은 기존 RNN과는 다르게 **위치 정보를 더해주는 포지셔널 인코딩**을 사용한다.

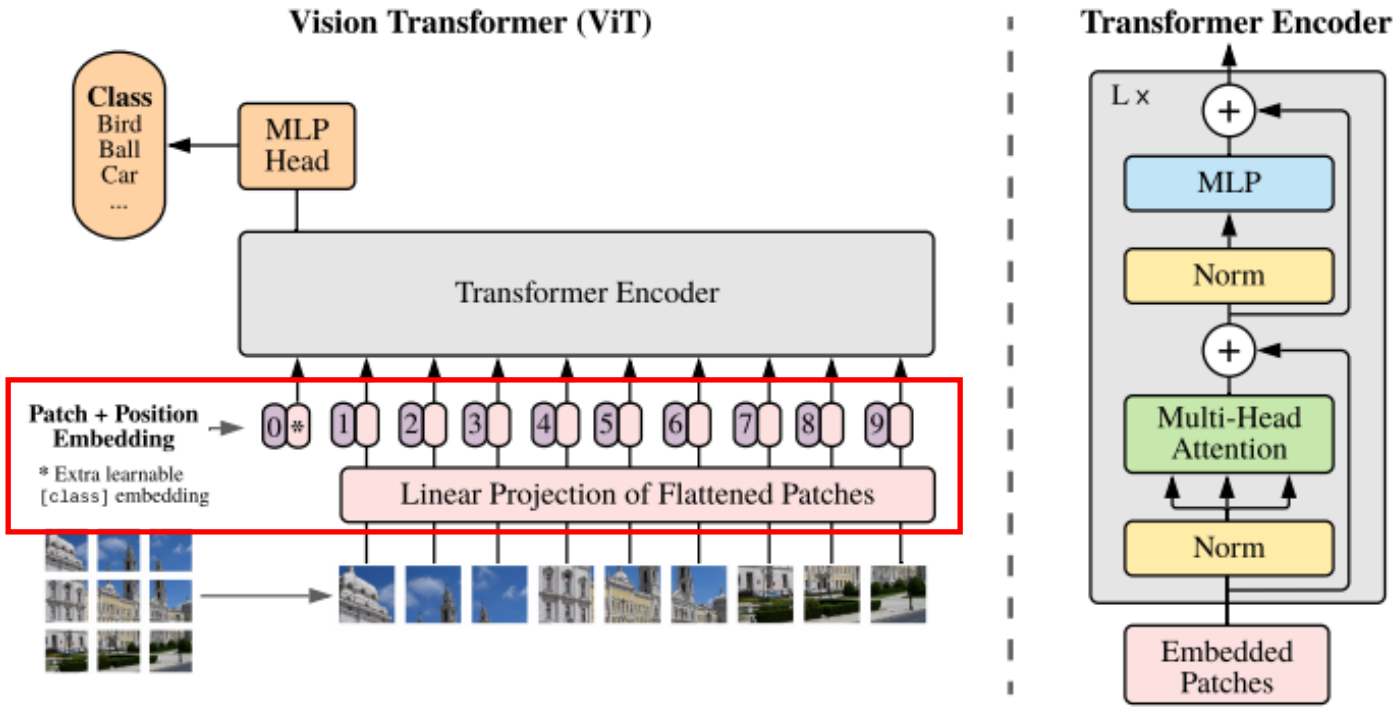
- Transformer는 연산이고, 확장성이 좋다. 특히 input sequence의 길이에 구애 받지 않는다.



<Transformer 구조>

Model

- 기존 Transformer와 최대한 같은 형태로 사용하기 위해 입력을 패치로 잘라 NLP의 토큰 처럼 사용한다. 그리고 위치에 대한 정보를 더해주기 위해 위치 임베딩을 사용한다.



<Vision Transformer 구조>

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 있었다.

	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>