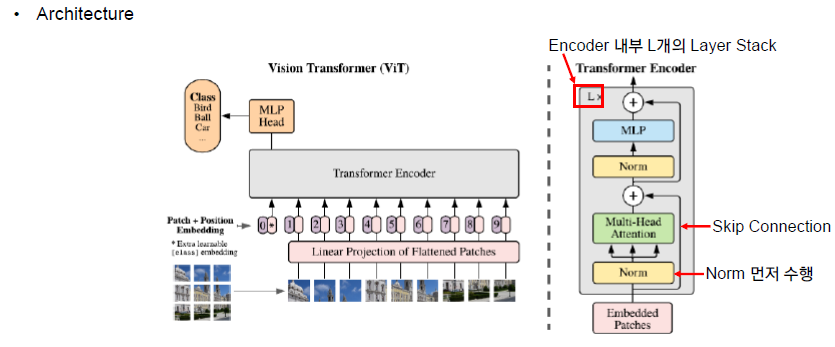
ViT : Vision Transformer

일반적으로 NLP에서 많이 사용되던 Transformer를 Vision Task에 적용한 논문.

기존의 제한적인 Attention 메커니즘에서 벗어나, CNN구조 대부분을 Transformer로 대체되었으며 inductive bias가 약한 대신 보다 통합적이고 일반적인 모델 구조라고 할 수 있음. inductive bias가 약하여 일반화 성능이 떨어지는 문제를 data-driven training, 즉 매우 큰 데이터셋에서의 사전학습을 통해 극복해냄

Transformer를 적용함으로 높은 연산 효율성(computational effect)와 확장성(scalability)를 갖게 되어 데이터셋과 모델 크기가 계속 커져도 모델 성능이 포화되지 않고 지속적으로 개선됨.

Vision 분야에 Transformer를 적용시킨 과정은 다음과 같음.



* 이미지 자체를 **standard Transformer**에 직접적으로 넣어주는 형태로 모델 구성
* 이를 위해 이미지를 패치 단위로 쪼개 "토큰화"시켜줌.
* 패치에 linear embedding 적용 후, 시퀀스로 만들어준 형태로 Transformer 입력으로 들어가게 됨.
* 중간 사이즈의 데이터셋에 학습을 시킬 경우, 별도의 강력한 regularization 없이 기존 ResNets 대비 훌륭한 성능을 보여주지는 못함.
* 이는 Transformer 구조 자체가 CNN 구조에 비해 inductive bias가 부족하여 많은 양의 데이터 없이는 일반화가 제대로 이루어지지 않았을 것으로 추정.
* 그러나 큰 사이즈의 데이터셋(1400만 ~ 3억장)에서 학습을 시킬 경우, 위의 구조적 한계 (lack of Inductive Bias)를 극복 가능함.
* ViT는 충분한 크기의 데이터셋(ImageNet-21k or JFT-300M)에서 사전학습 후, 보다 적은 데이터셋(ImageNet, ImageNet-Real, CIFAR-100, VTAB)을 가진 task에 전이 학습을 시킬 때 좋은 성능을 보여주었음.