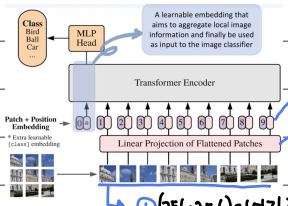
Vision Transformer ViT

- 기존의 Transformer은 아래 그림처럼 Input에는 문장벡터값이 들어갔다.

- 하지만 ViT로 넘어오면서 여러 부분이 바뀐다.

① Input



③ 클래스 토큰 추가 → (N+1, D)

Position Encoding은 (N+1, D) 사이즈를 만들어 그대로 더해준다. → (N+1, D)

② [768, 모델파라미터 D] 와 연산을 통해 [256, D]로 만든다.

→ (N, D)

① (256·256) 이미지를 (16·16)으로 나눠서 256개의 패치생성 후에

1개의 패치당 RGB를 고려해 768크기의 벡터 생성한다.

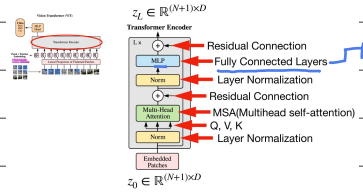
[256, 768]

전체 이미지에 대한 정보를 종합하고 분류 작업을 수행하는데 중요하게 작용한다.

각 위치 정보가 각 패치 임베딩에 포함된다.

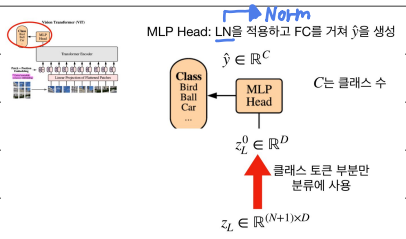
② Transformer Encoder

- L번 반복하기 위해 입력과 출력의 크기가 같도록 유지한다.
- 기존의 Transformer Encoder에서는 Multi-Head 하고 Norm을 하는데 순서가 바뀌어있다.



FC층 2개, GELU사용

- 입력과 입력의 누적정규분포의 곱을 사용한 형태이다.
- 모든곳에 사용가능, 단조 증가 함수 아니라 Activation Function 함수로 사용가능
- 입력값 x가 다른 입력에 비해 얼마나 큰지 비율로 값이 조정되어 확률적 해석가능



이렇게나오면 목적에 맞게 활성화함수를 정하고 모델을 학습한다.