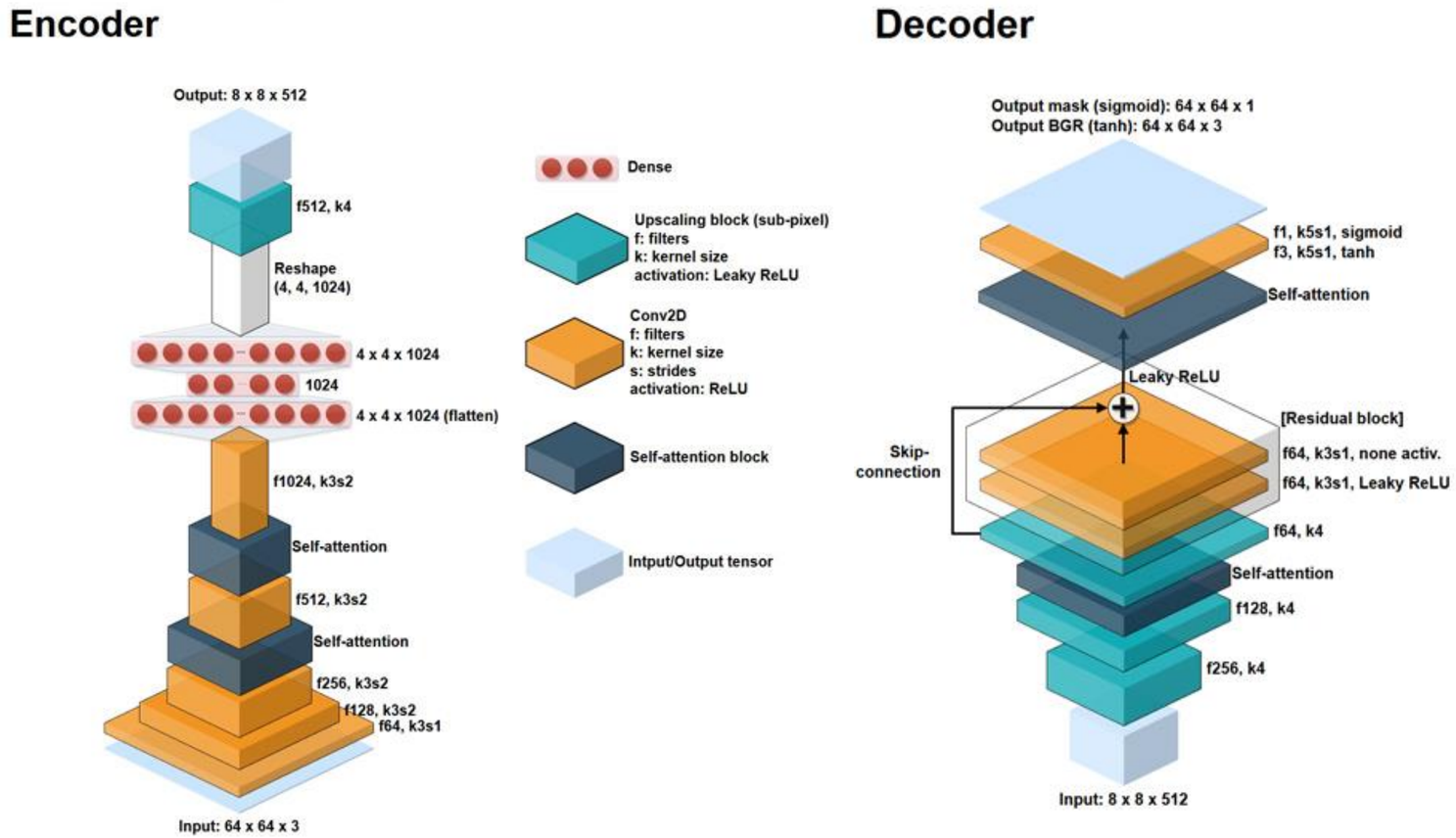
Vision Transformer

생성형 AI Study
Luddite

발표자 : 김채아

Deepfake?

- Encoder-Decoder 구조
- Sequence data
- Self-attention, skip-connection



Faceswap-GAN

Transformer

Sequential data를
처리하고 인코딩 하는 모델

- Translation
- Image classification (ViT)
- Image detection (DETR)
- DALL-E, Deepfake 등 transformer 구조인 self-attention 활용

Transformer 구조

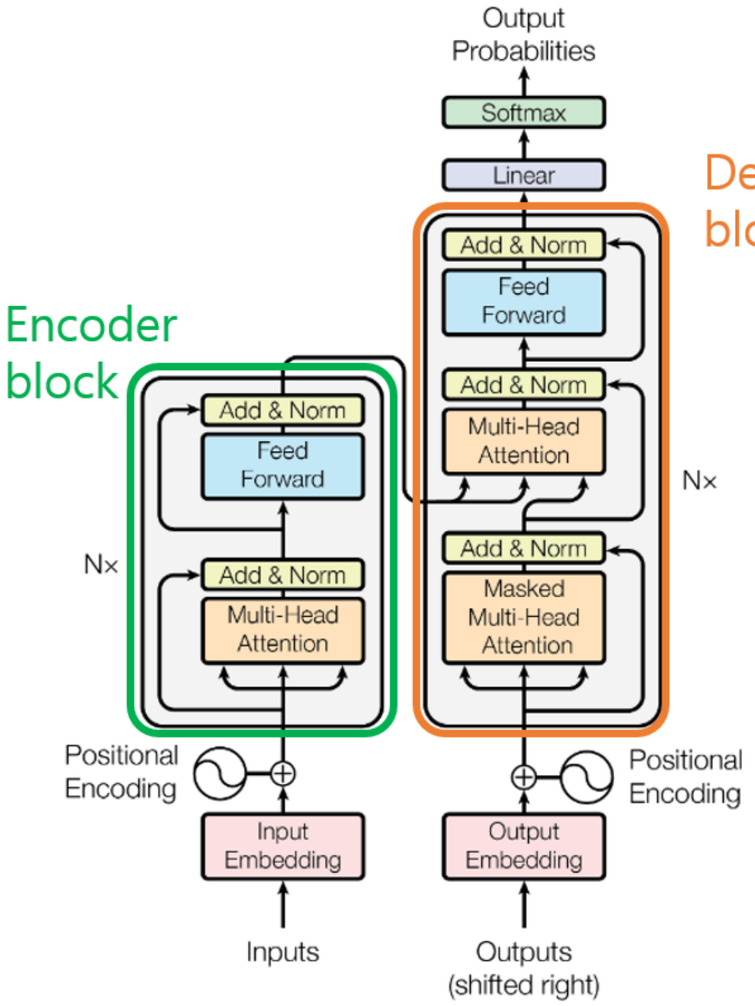
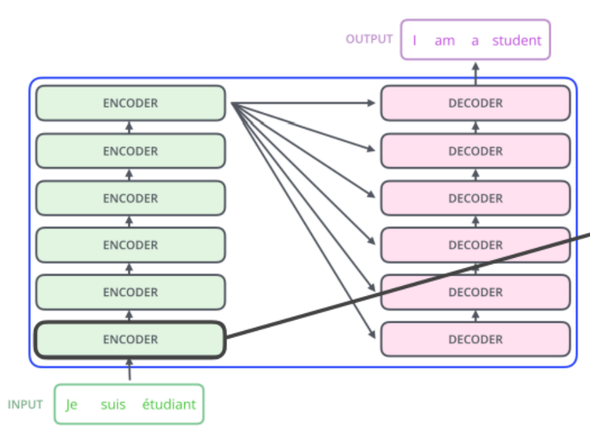
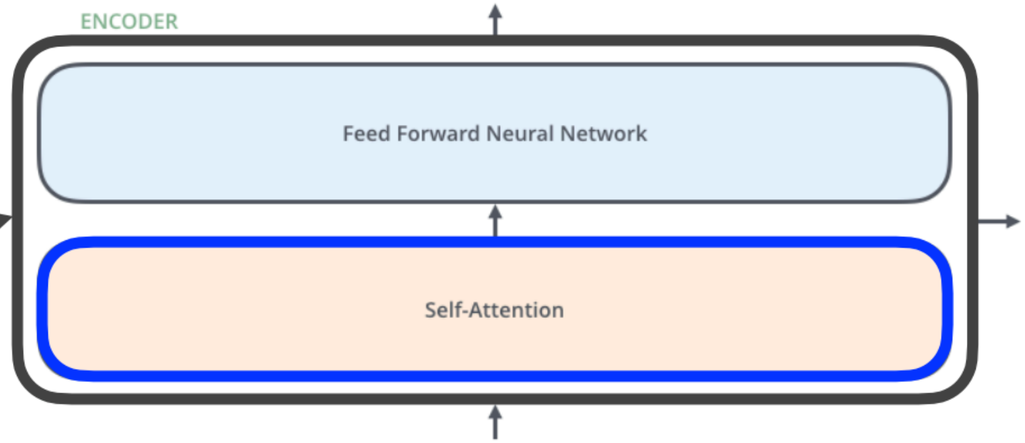


Figure 1: The Transformer - model architecture.

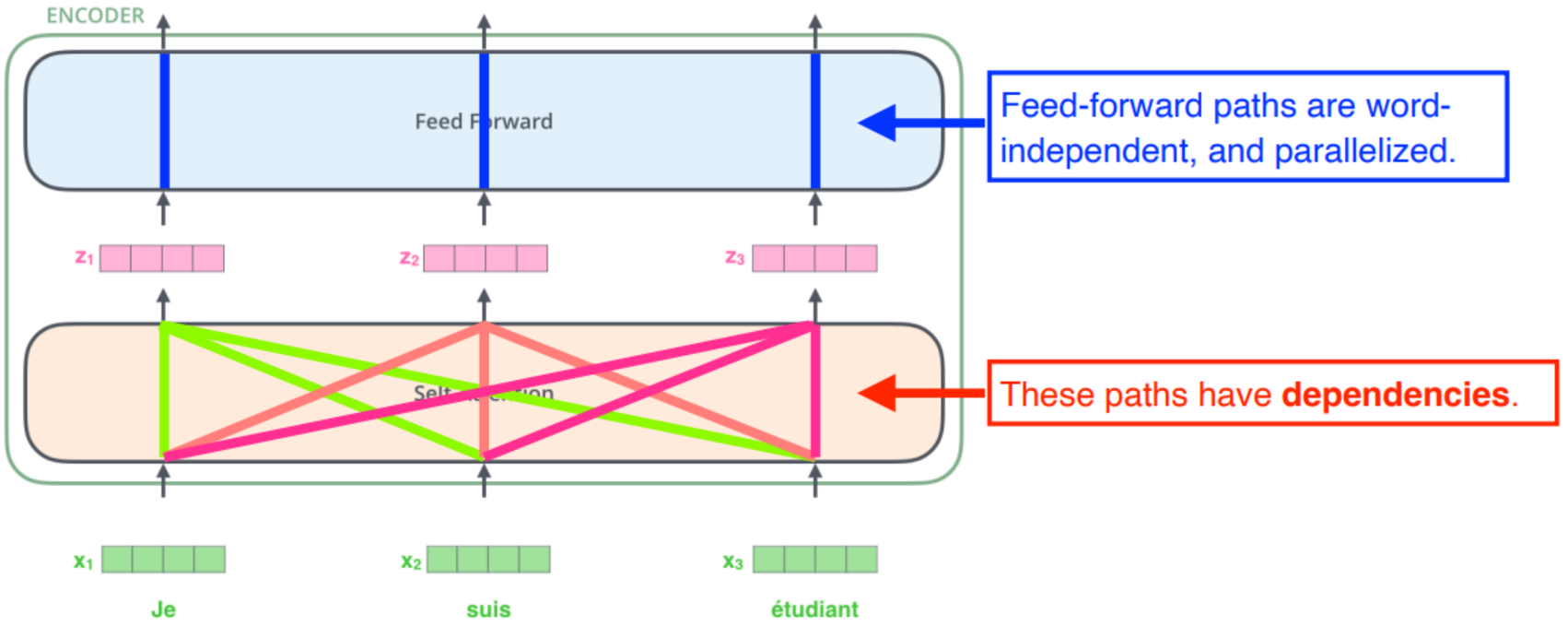




Self-Attention!

https://arxiv.org/abs/1706.03762

Self-attention

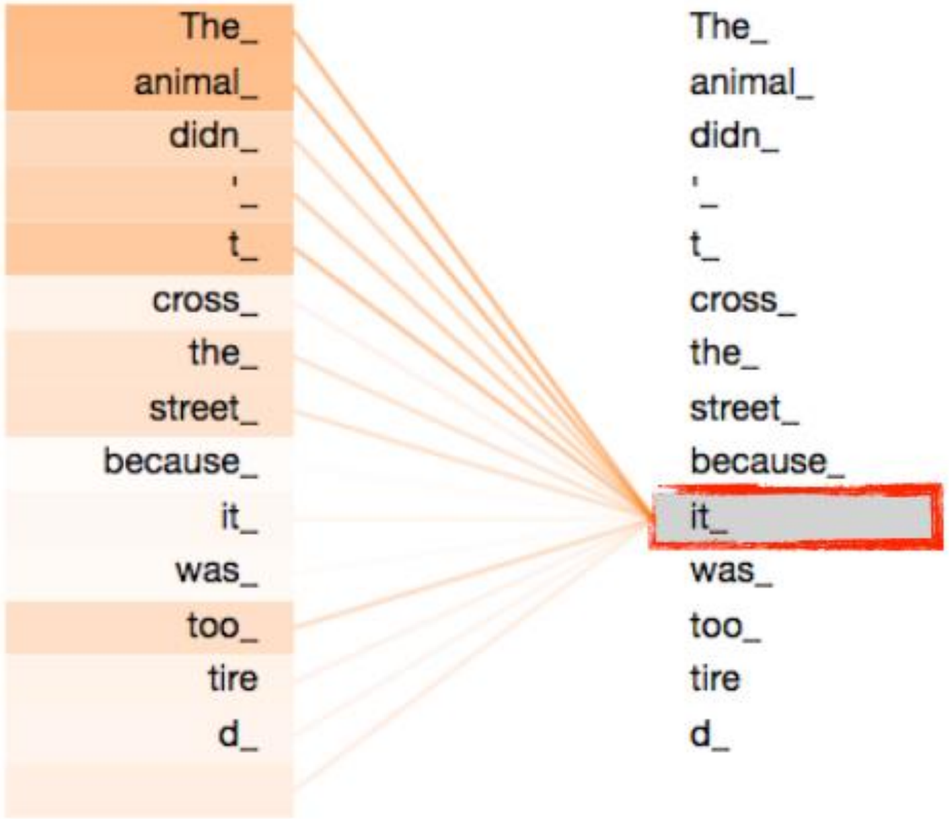


n개의 입력 x가 주어지고 z벡터를 찾는데,
i번째 x벡터를 zi로 바꿀 때 나머지 n-1개의 x벡터를 같이 고려

http://jalammar.github.io/illustrated-transformer/

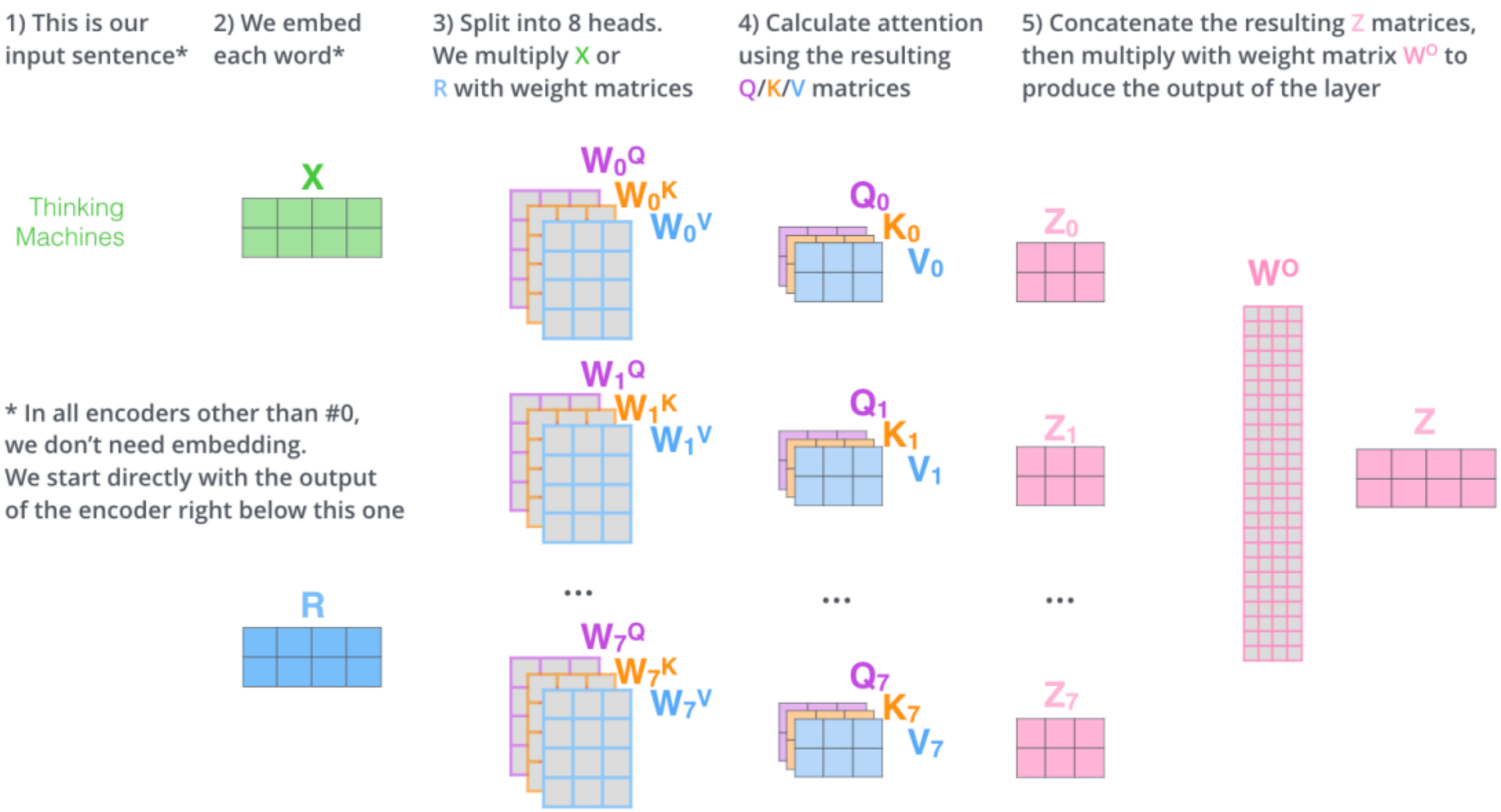
Self-attention

The animal didn't cross the street because it was too tired.



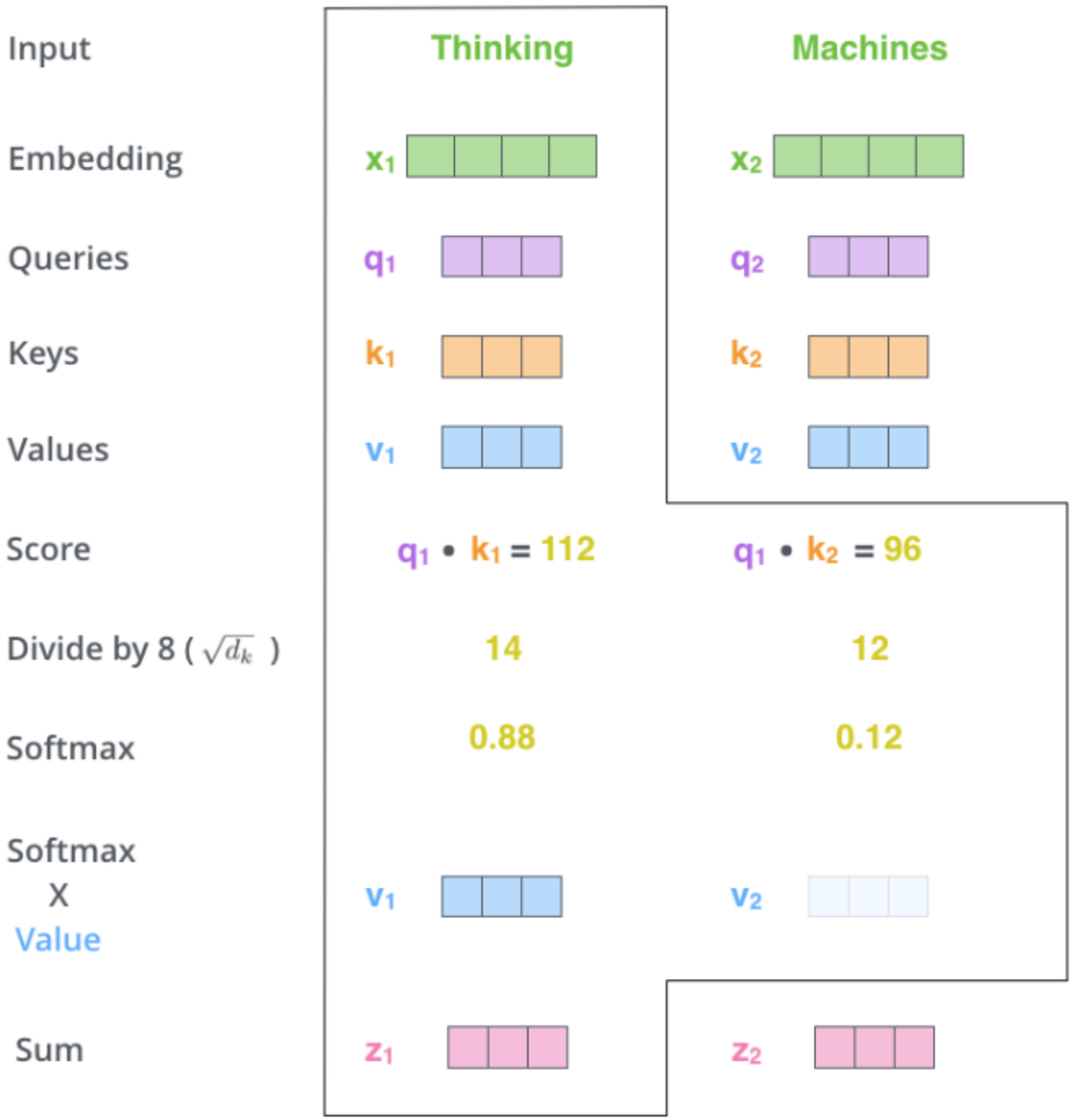
http://jalammar.github.io/illustrated-transformer/

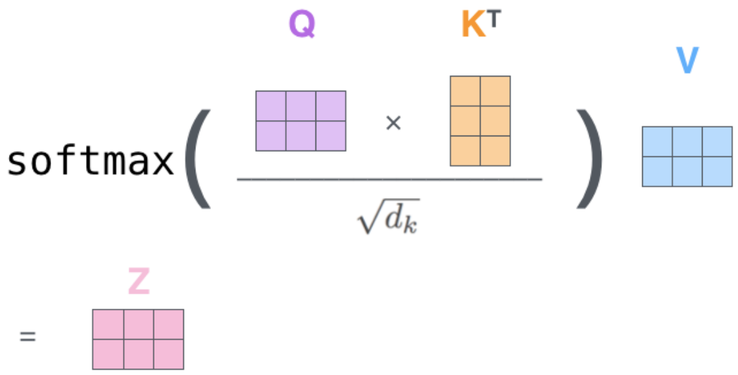
Self-attention



http://jalammar.github.io/illustrated-transformer/

Self-attention





http://jalammar.github.io/illustrated-transformer/

Vision Transformer

NLP task에서 SOTA를 찍었던
모델인 Transformer를
CV분야에 적용한 방법론

- CNN보다 inductive bias가 부족

  데이터가 적으면 ResNet보다 정확도 낮음

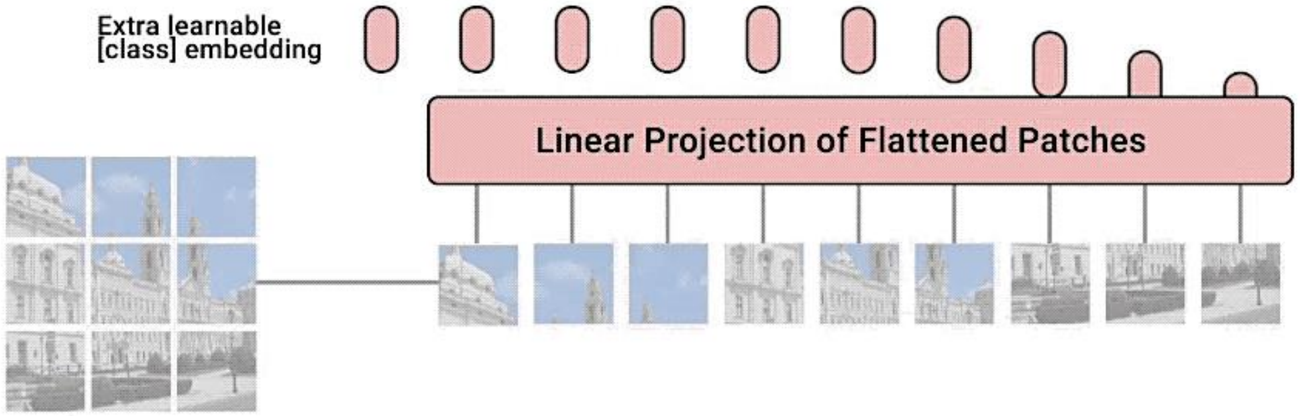
  -> 더 많은 데이터를 요구

- Large scale dataset 활용

  CNN보다 확연히 높은 성능 보임

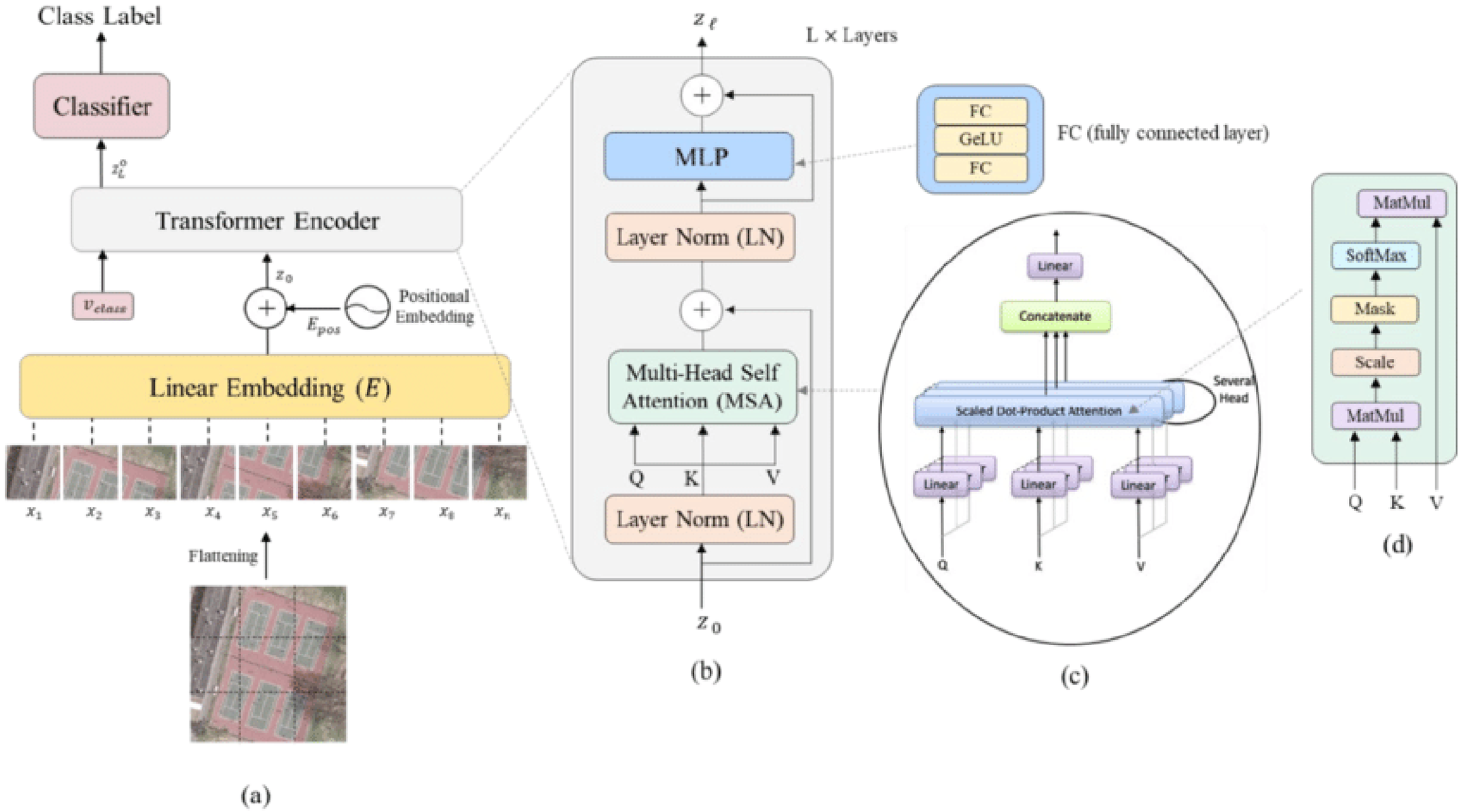
Method of ViT

(1) 이미지 patch로 단위화
(2) Linear Projection of Flattened Patches
(3) Class token 추가, positional embedding
(4) Transformer Encoder 통과
(5) MLP Head 통과
(6) Class 분류



https://velog.io/@tm011899/%EB%84%A4%EC%9D%B4%EB%B2%84-%EB%B6%80%EC%8A%A4%ED%8A%B8%EC%BA%A0%ED%94%84-5%EA%B8%B0-19%EC%9D%BC%EC%B0%A8

Method of ViT



감사합니다