VIT

RNN은 오랜시간동안 NLP분야에서 사용됨

하지만 현재는 Transformer 중심으로 이루어지고 있음.

Transformer in Computer Vision

CNN에 Transformer를 붙여 사용

↓

Transformer만을 이용한 모델링

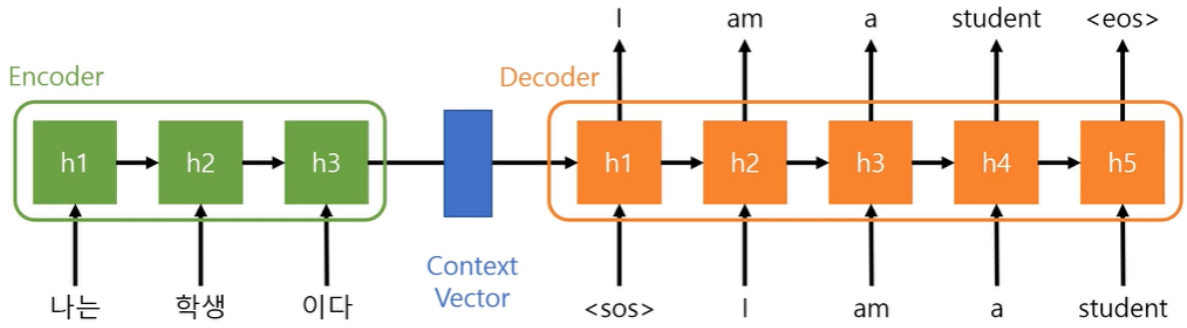
Transformer란?

Attention만을 활용한 모델 구축(self attention)

*Attention (Seq2Seq)

Seq2Seq - 문장을 입력으로 받아 문장을 출력하는 모델 / 기계번역에 주로 사용

Context Vector - Decoder에게 전달되는 입력 문장의 정보

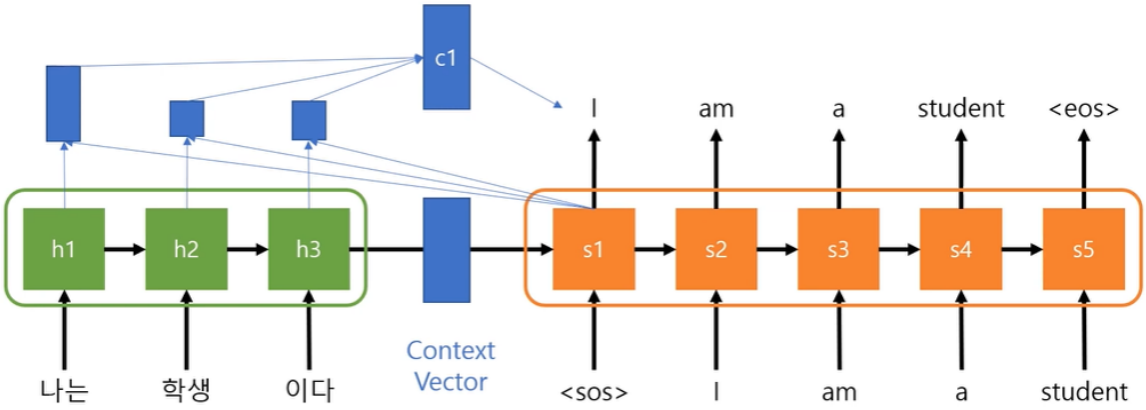


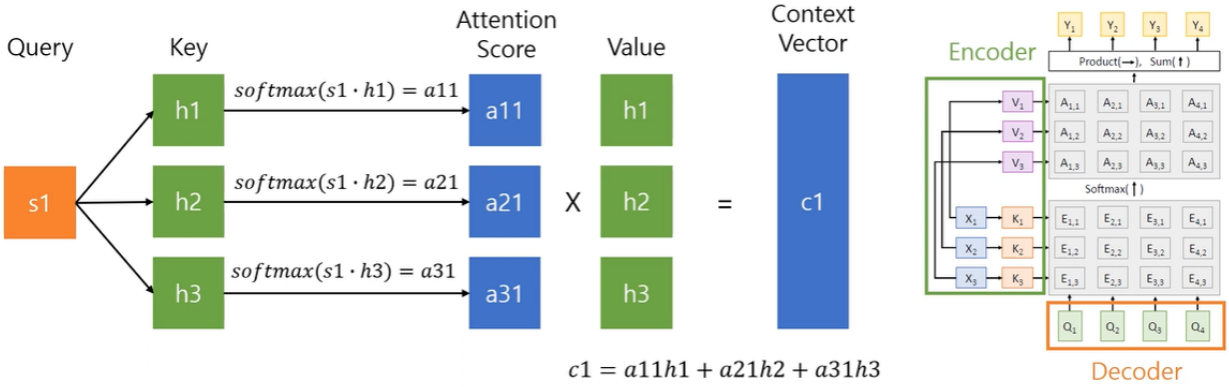
***Context Vector의 크기가 제한적이기 때문에 입력 문장의 모든 정보 전달이 불가**

↓

Seq2seq with Attention

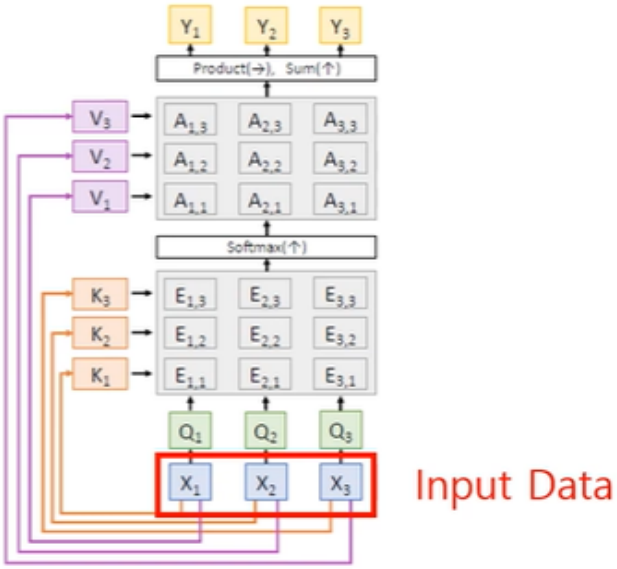
- Decoder가 특정 시점 단어를 출력할 때 encoder 정보 중 연관성이 있는 정보를 직접 선택





*Encoder와 Decoder 부분의 output 간의 상관관계를 통해 특징을 추출

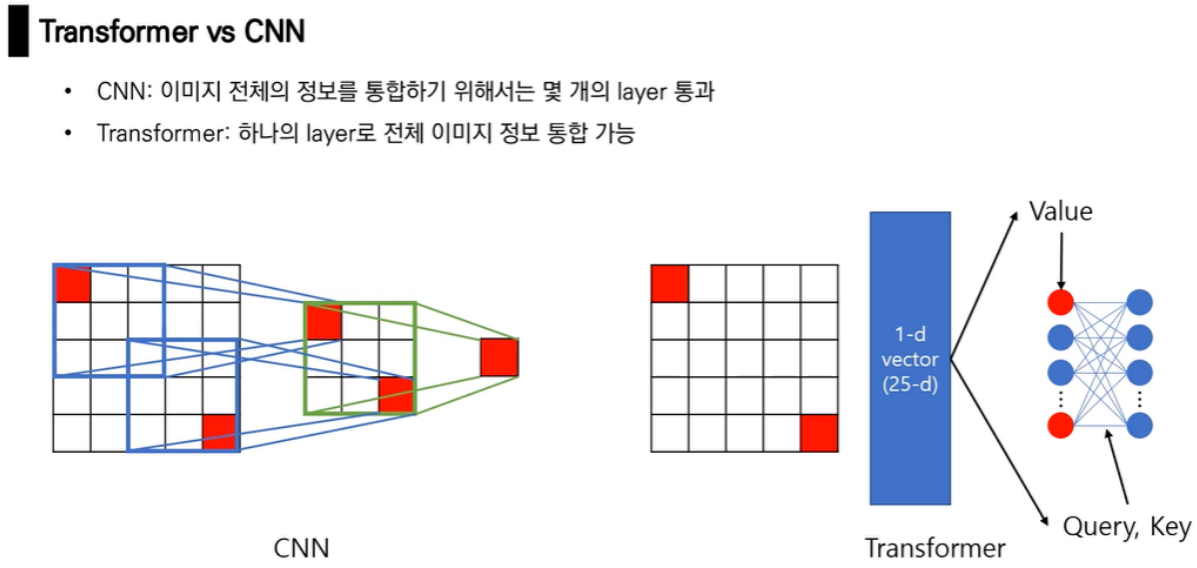
Self Attention



데이터 내의 상관관계를 바탕으로 특징을 추출해 나가는 과정

CNN vs Transformer

1.Correlation



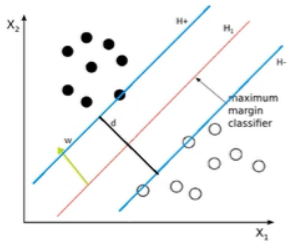
빨간네모 간의 상관관계를 파악해야 한다고 가정

CNN - Convolutional Layer가 두 번 통과해야 상관관계 파악이 가능

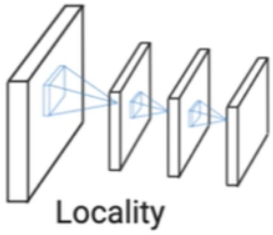
Transformer - 1차원의 벡터로 풀어 Self Attention을 적용

2.Inductive Bias

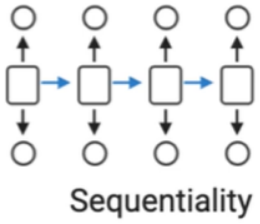
새로운 데이터에 좋은 성능을 내기 위해 모델에 사전적으로 주어지는 가정



SVM



CNN



RNN

SVM - Margin 최대화

CNN - 지역적인 정보 추출(Convolution Filter)

RNN - 순차적인 정보

CNN - 지역적인 특성을 유지해야 함, 학습 후 Weight 고정

Transformer - 1차원 벡터로 만든 후 Self Attention진행(지역적 정보 x)

Weight가 input에 따라 유동적으로 변함

- TF : Inductive Bias ↓, 자유도 ↑