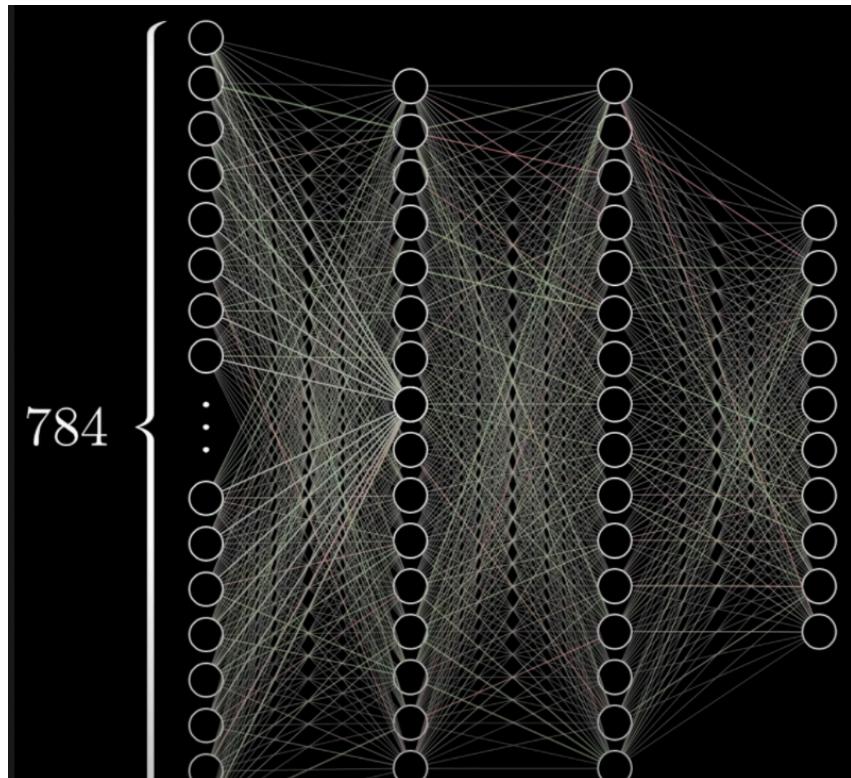


RNN → Encoder-Decoder → Attention → Transformer → BERT

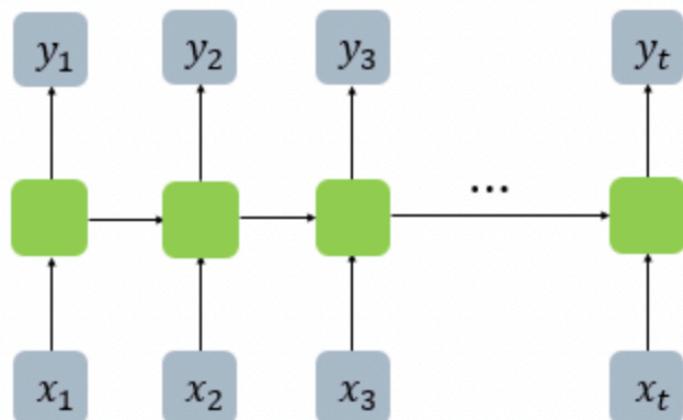
Vanilla FCN

- Fixed Input Length

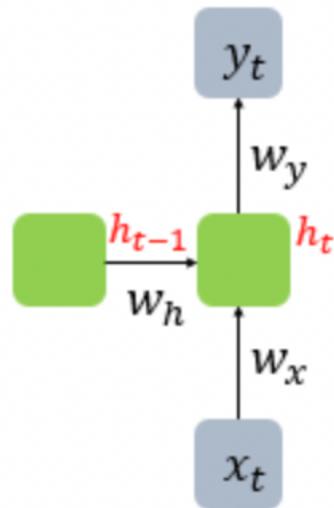


RNN (Recurrent Neural Network)

- One to Many
- Many to One
- Many to Many
- 입력의 길이가 다양해서 텍스트 (자연어) 처리에 유용



- 은닉상태 (hidden state) : 현재 시점 t에서 메모리 셀이 가지는 값



RNN의 문제점

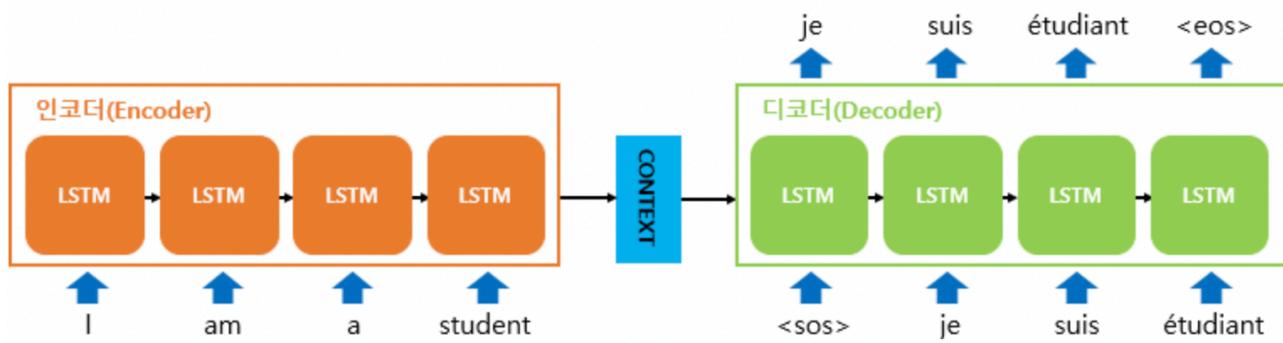
- 장기의존성 (Long-Term Dependency): time-step이 길어질수록 앞쪽의 정보가 소실
- LSTM (느림)

Encoder - Decoder

2개의 RNN을 사용해서 각각 Encoder, Decoder라고 명명하고 이어붙인 형태 (Concat)

seq2seq

- 자연어 처리에서 Encoder - Decoder 구조로 사용되는 대표적인 형태



입력 -> 인코더 -> Context Vector -> 디코더 -> 출력

Attention

- 고정된 크기를 가진 Context Vector에 모든 정보를 압축하니 필연적인 정보의 손실이 발생
- Attention은 이 점을 보완하기 위해 고안된 방법

“어텐션은 매 time-step마다 인코더의 전체 입력을 다시 한번 본다.”

Query (Q) : t 시점의 디코더 셀에서의 은닉 상태

Key (K) : 모든 시점의 인코더 셀의 은닉 상태들

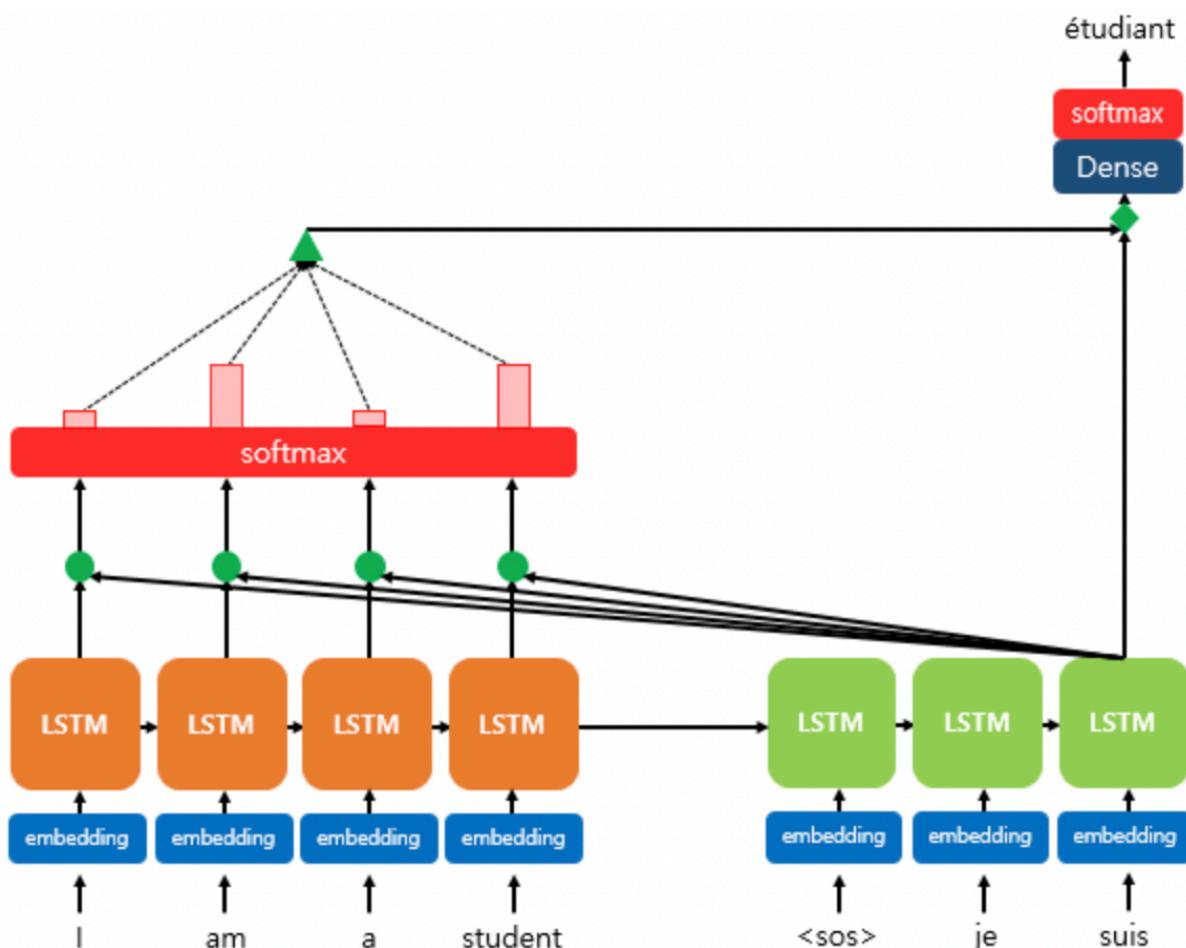
Value (V) : 모든 시점의 인코더 셀의 은닉 상태들

인코더의 전체 입력을 본다 —>

1. Query에 대해 모든 Key와의 유사도를 계산 (Attention Score)
2. 소프트맥스 함수를 취해서 합이 1인 확률 분포를 얻는다 (Attention Weight)
3. Attention Weight와 Value를 가중합(weighted sum) —> Attention Value (Context Vector)

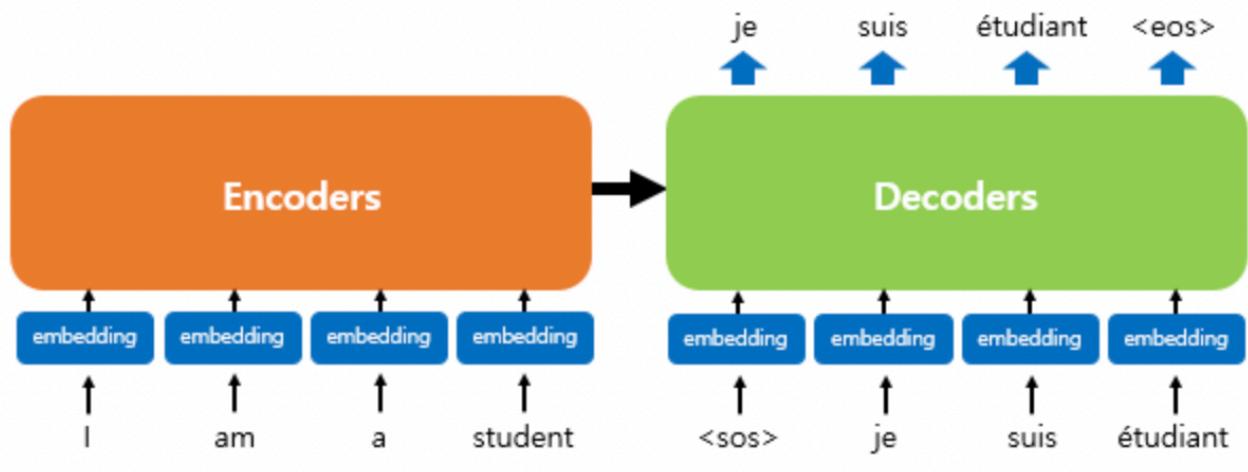
Attention Value와 t시점의 디코더의 은닉 상태를 연결

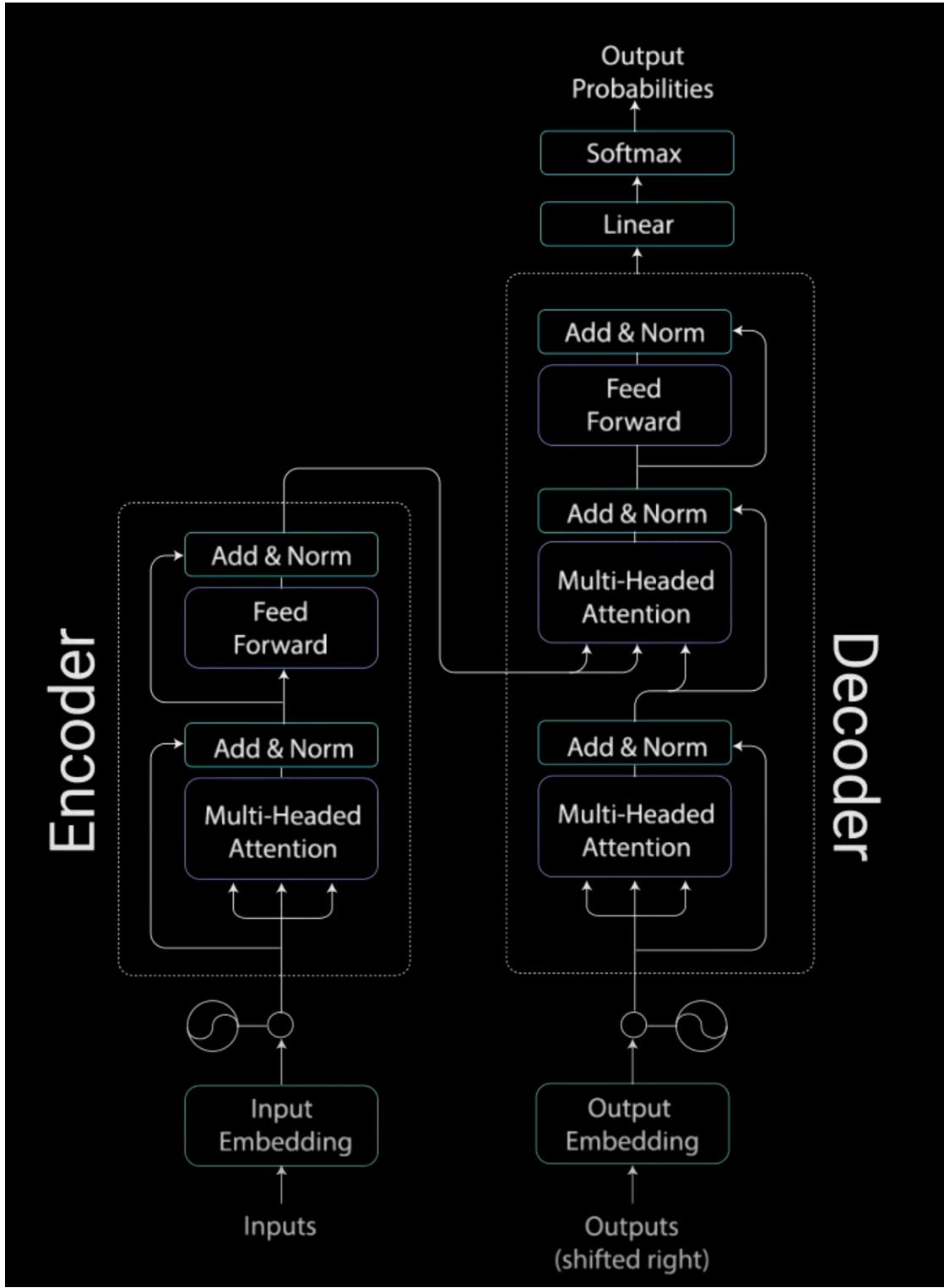
—> 디코더의 값에 인코더의 정보가 반영됨 (핵심 아이디어)



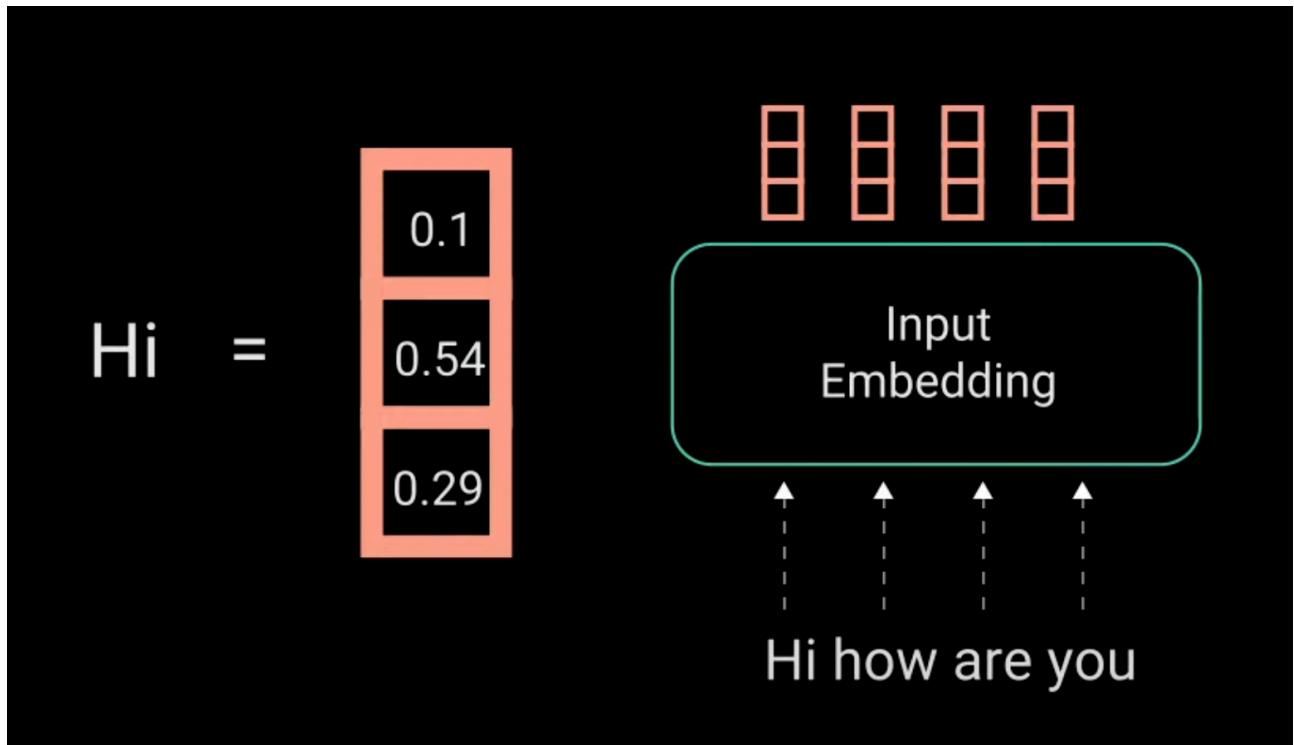
Transformer

- 순수하게 어텐션만으로 Encoder - Decoder 구조를 쌓아올려 만듬
- BERT, GPT 등 자연어 처리의 힘쓸고 있는 모델들의 근간이 되는 구조



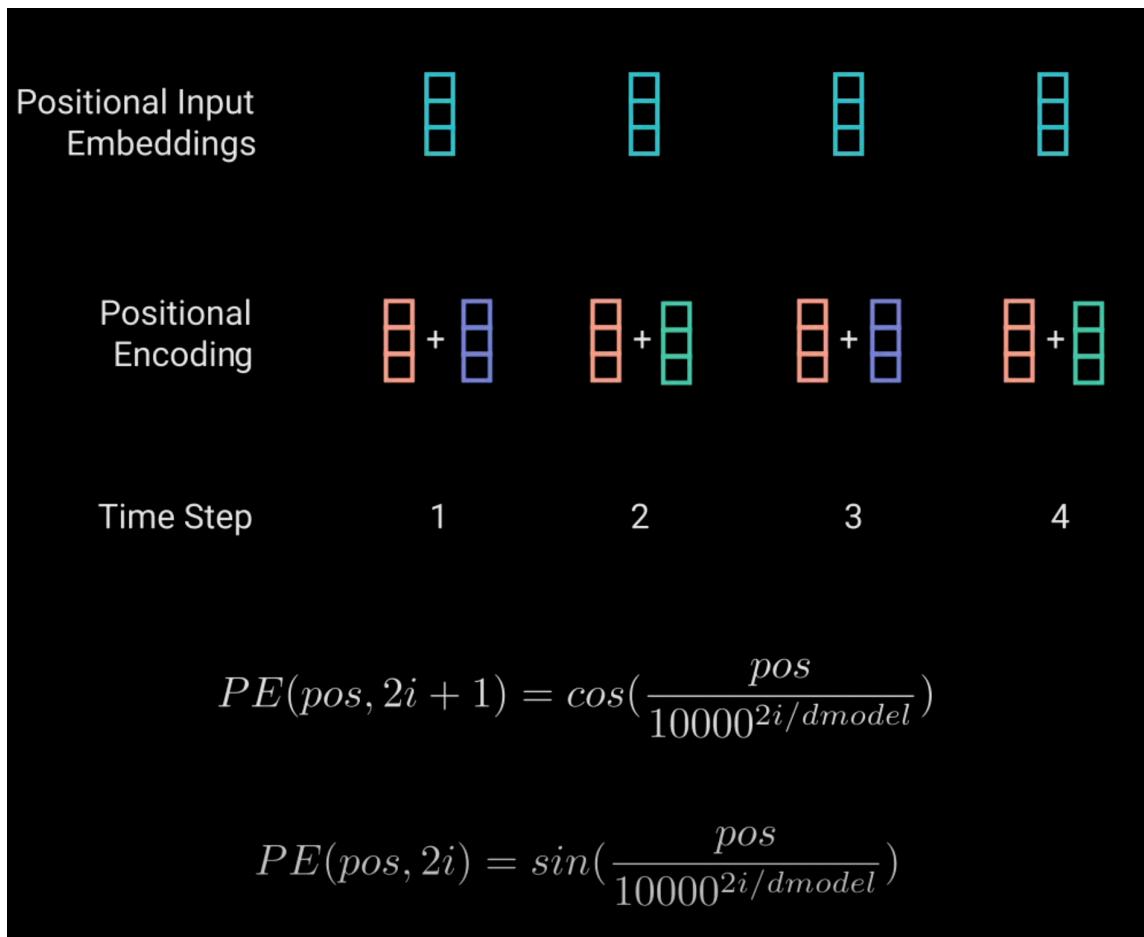


Input Embedding



Positional Encoding (위치 인코딩)

- RNN의 장점은 입력을 순차적으로 받아서 처리하므로 단어의 위치정보를 알 수 있었음
- 트랜스포머는 입력을 동시에 처리하므로, 위치 정보를 인위적으로 추가해줘야 함
- Sine(짝수), Cosine(홀수) 함수를 사용
- 같은 단어라도 문장 내의 위치에 따라서 트랜스포머의 입력으로 들어가는 임베딩 벡터의 값이 다름

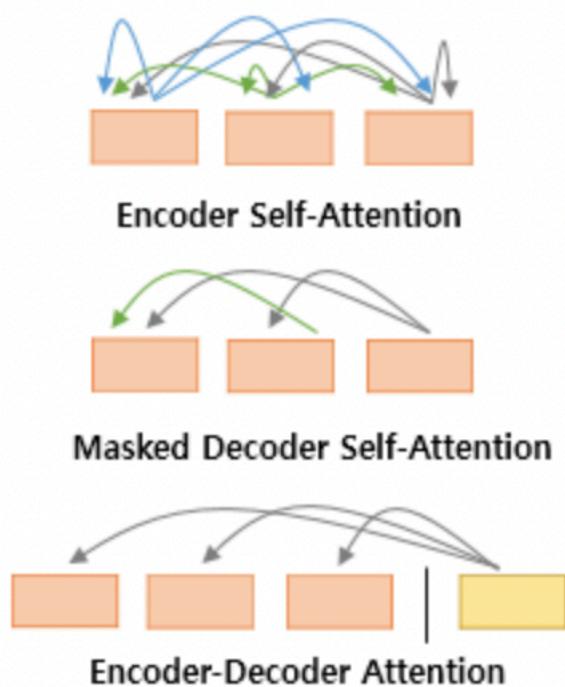


Attentions

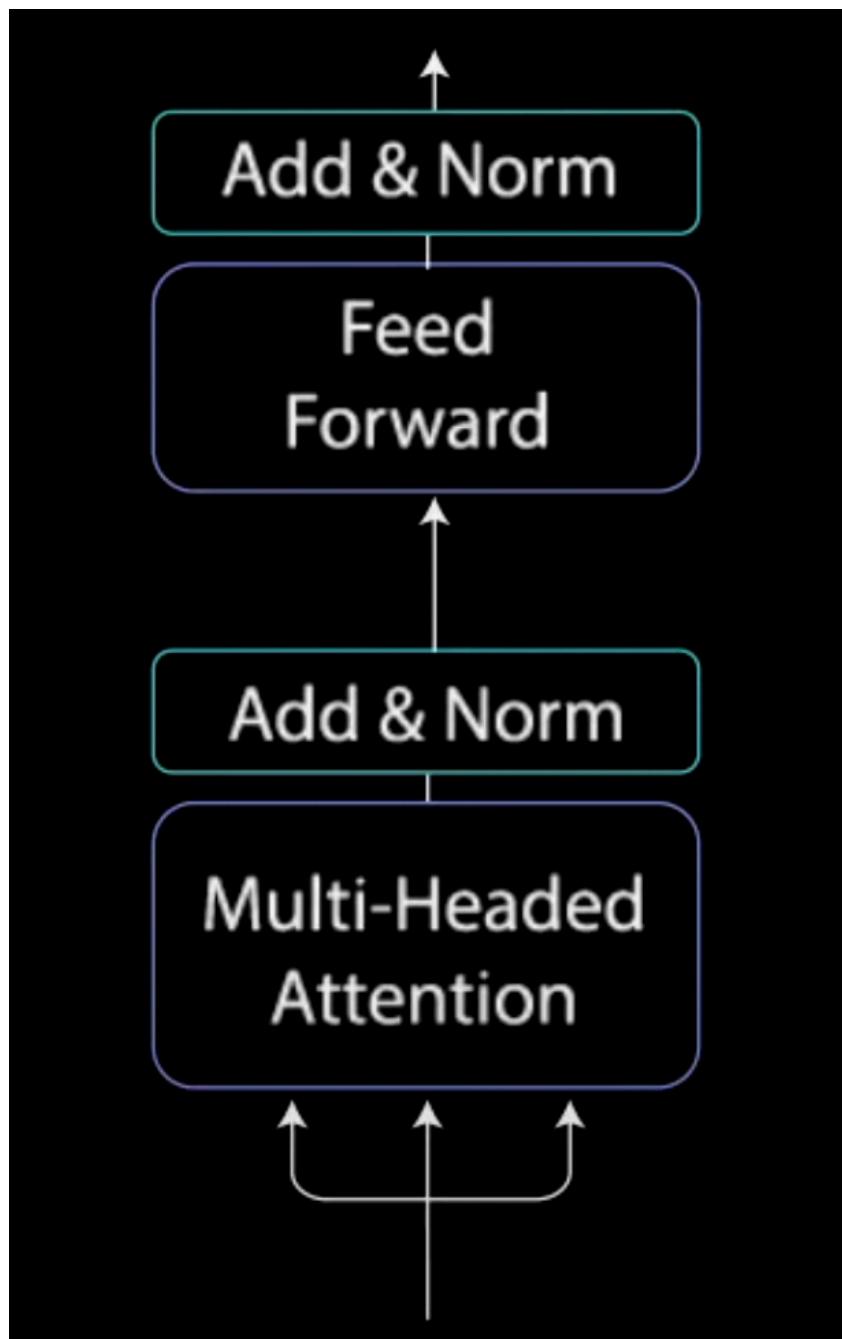
“트랜스포머에서는 3가지 종류의 어텐션이 사용됨”

- Encoder Self-Attention
- Masked Decoder Self-Attention
- Encoder-Decoder Attention

셀프어텐션이란 Q , K , V 의 값의 출처가 동일하다는 의미

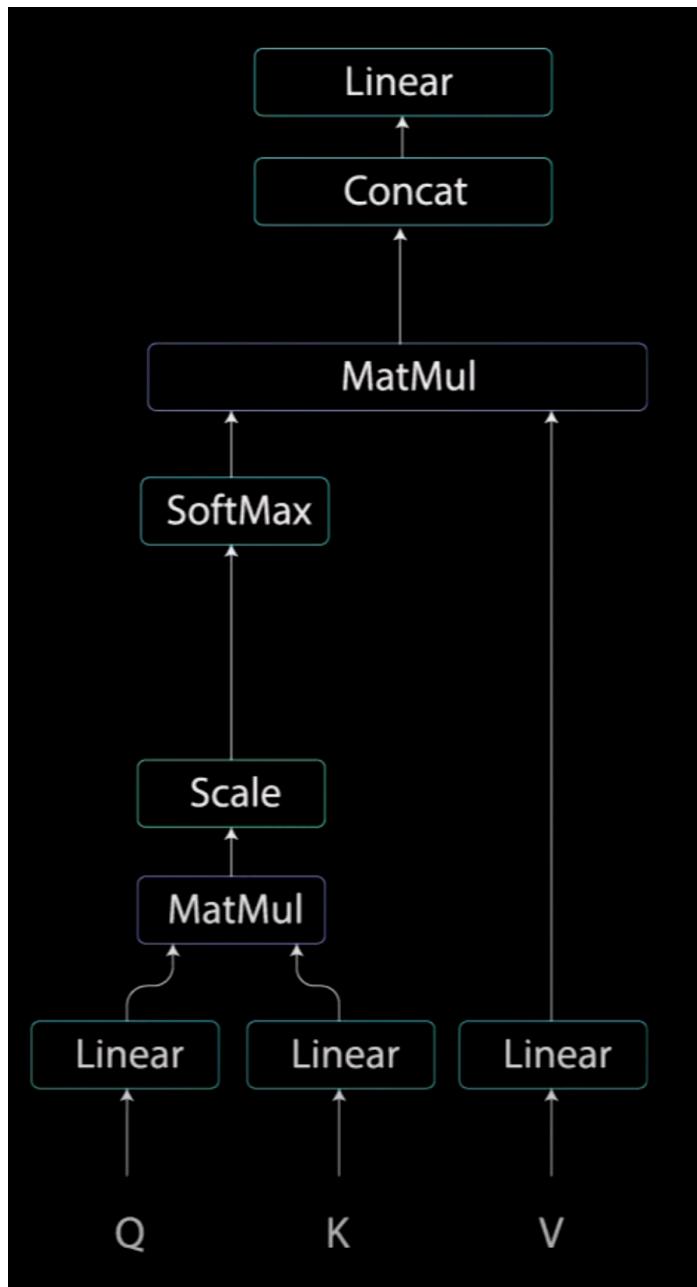


Encoder



Encoder Self-Attention

- 입력된 단어 벡터로부터 Q, K, V 벡터를 만듬



Multi-Head Attention

- Multi-Head : 셀프 어텐션을 병렬적으로 사용 (Q, K, V 를 더 쪼慨)
- 예) 8차원벡터의 Q, K, V, 2 heads이면 각 Q, K, V는 4차원벡터로 분리되어 입력

Add & Norm

- 학습에 도움을 주기 위한 기법

- Add

: Residual Connection (잔차 연결)

출력값에 입력값을 또 한번 더해주는 것

- Norm

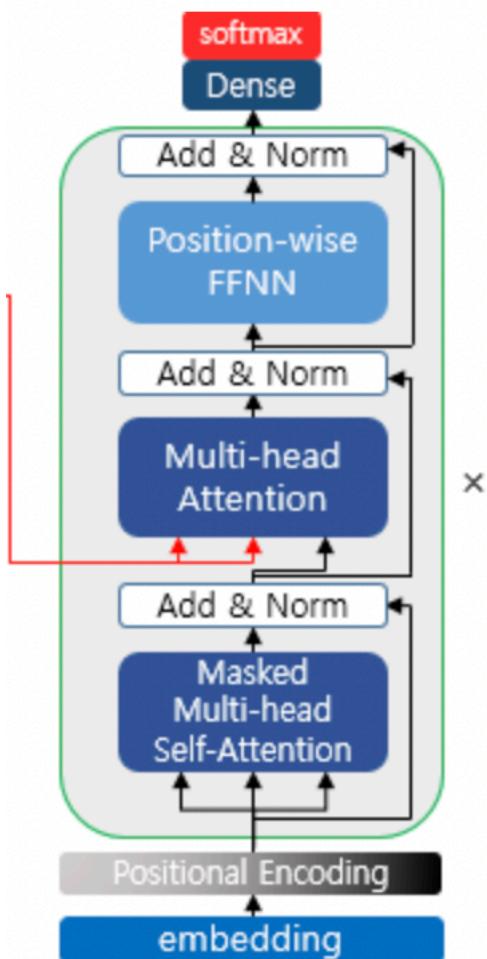
: Layer Normalization (층 정규화)

Feed Forward

- 완전 연결 신경망 (FCN)

Decoder

- 디코더는 마지막 인코더의 출력값을 입력으로 전달 받음
- 트랜스포머는 Teacher Forcing 을 사용하여 훈련
- Teacher Forcing 이란 t 시점의 출력이 t+1시점의 출력으로 사용되는 모델에서 모델의 예측 값을 t+1값에 사용하는 것이 아닌 실제 정답을 t+1값으로 사용하는 방법



Masked Decoder Self-Attention

- 현재 시점의 단어를 예측하고자 할 때, 입력 문장 행렬로부터 미래 시점의 단어까지도 참고할 수 있는 현상이 발생을 막기 위해 미래 시점의 단어를 masking하고 셀프어텐션을 진행

- Query에 대해 모든 Key와의 유사도를 계산하는 것 대신 현재 시점 까지의 Key와의 유사도만 계산

Encoder-Decoder Attention

- Query는 디코더, Key, Value는 인코더에 가져와서 어텐션 계산

Multi-head Attention

Add & Norm

BERT

- 트랜스포머의 인코더만 쌓아 올린 형태