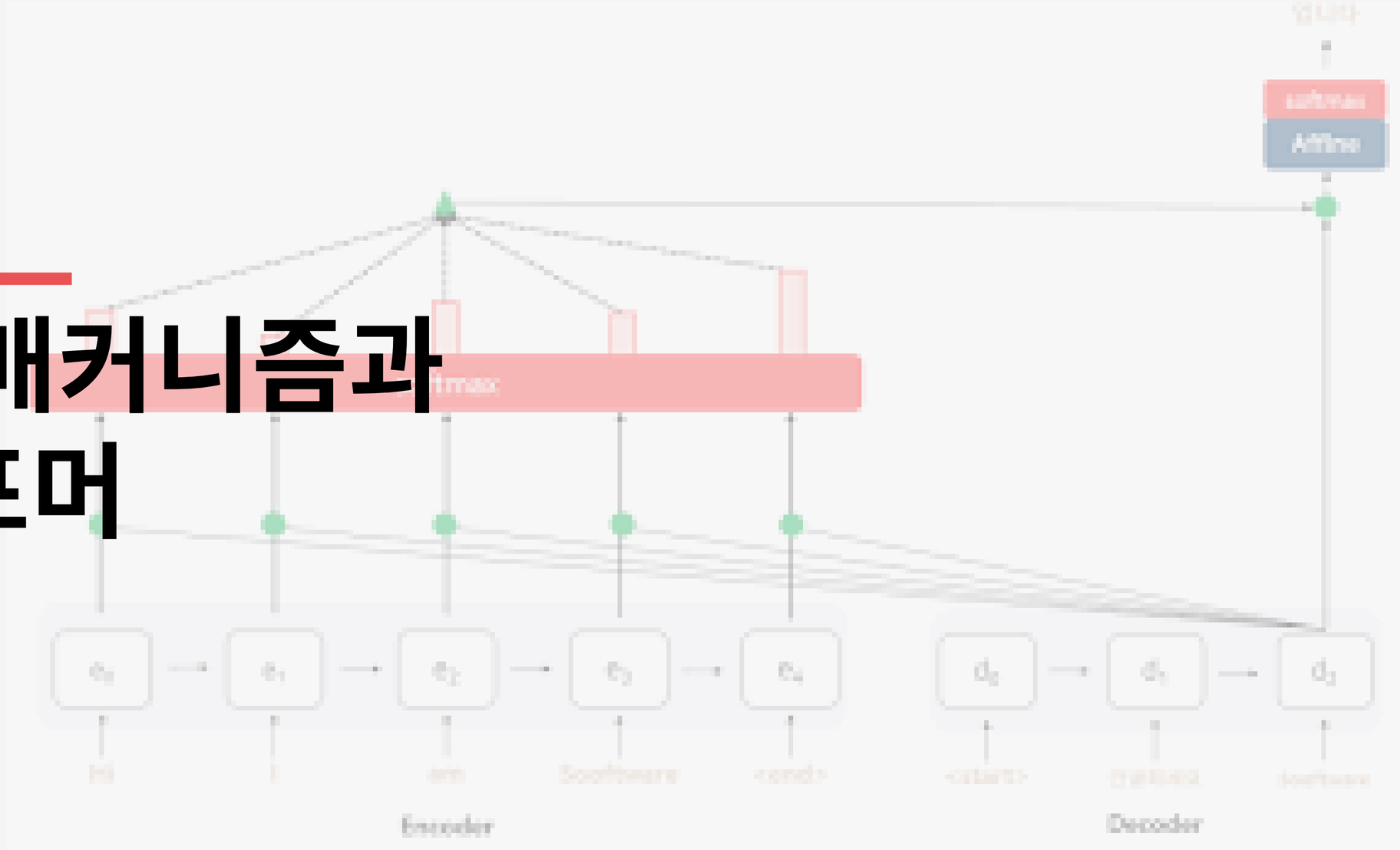
어텐션 매커니즘과 트랜스포머

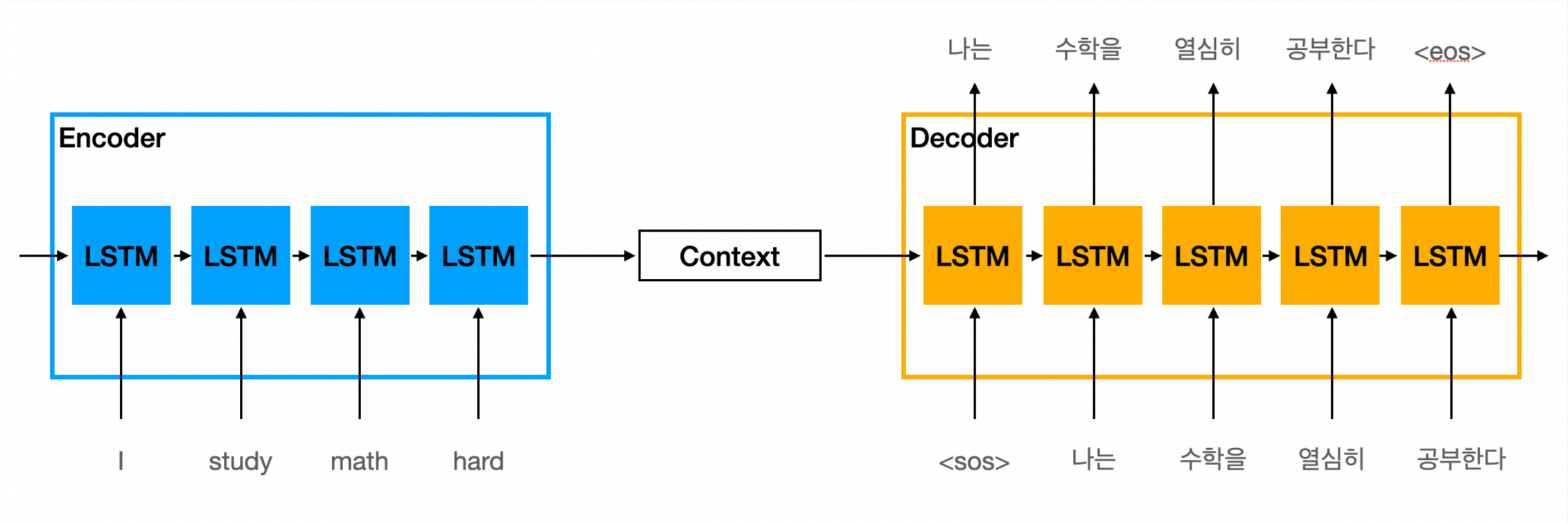
인공지능 명예학회



Contents

- seq2seq의 단점
- 어텐션 메커니즘
- 핵심 원리
- 트랜스포머
- 트랜스포머를 이용한 다양한 모델
- 코드 실습

기존 seq2seq의 단점



기존 seq2seq의 단점

기울기 손실

RNN의 고질적인 문제로,
역전파 과정에서 입력층으로
갈 수록 기울기가 줄어든다.

정보 손실

하나의 벡터에
맥락을 압축하다보니
정보가 손실됨

어텐션 메커니즘

디코더에서 출력 단어를 예측하는 시점마다, 인코더에서의 전체 입력 문장을 다시 한번 참고한다. 전체 입력 문장을 동일한 비율로 참고하는 것이 아니라, 해당 시점에서 예측해야 할 단어와 연관이 있는 부분을 더 집중(attention)해서 보게 된다.

이 뒤의 어텐션 메커니즘에 대한 서술은 인코더-디코더 구조에서 디코더 부분에 어텐션 메커니즘을 추가한 경우이다.

어텐션 메커니즘

Attention(Q, K, V) = Attention Value

Q = Query : t 시점의 디코더 셀에서의 은닉 상태

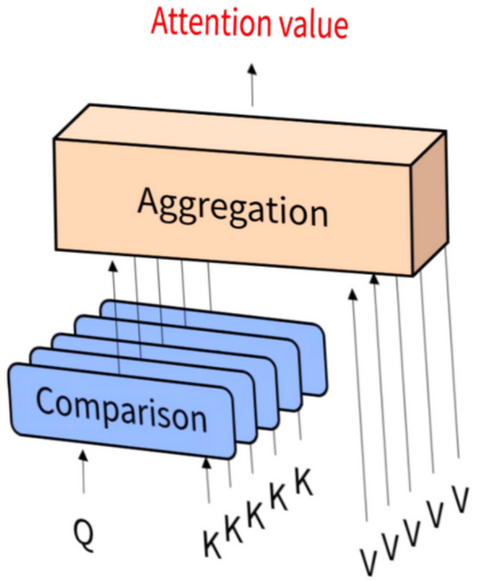
K = Keys : 모든 시점의 인코더 셀의 은닉 상태들

V = Values : 모든 시점의 인코더 셀의 은닉 상태들

주어진 '쿼리(Query)'에 대해서 모든 '키(Key)'와의 유사도를 각각 구한다.
그리고 구해낸 이 유사도를 키와 맵핑되어있는 각각의 '값(Value)'에 반영한다.
유사도가 반영된 '값(Value)'을 모두 더해서 리턴한다.

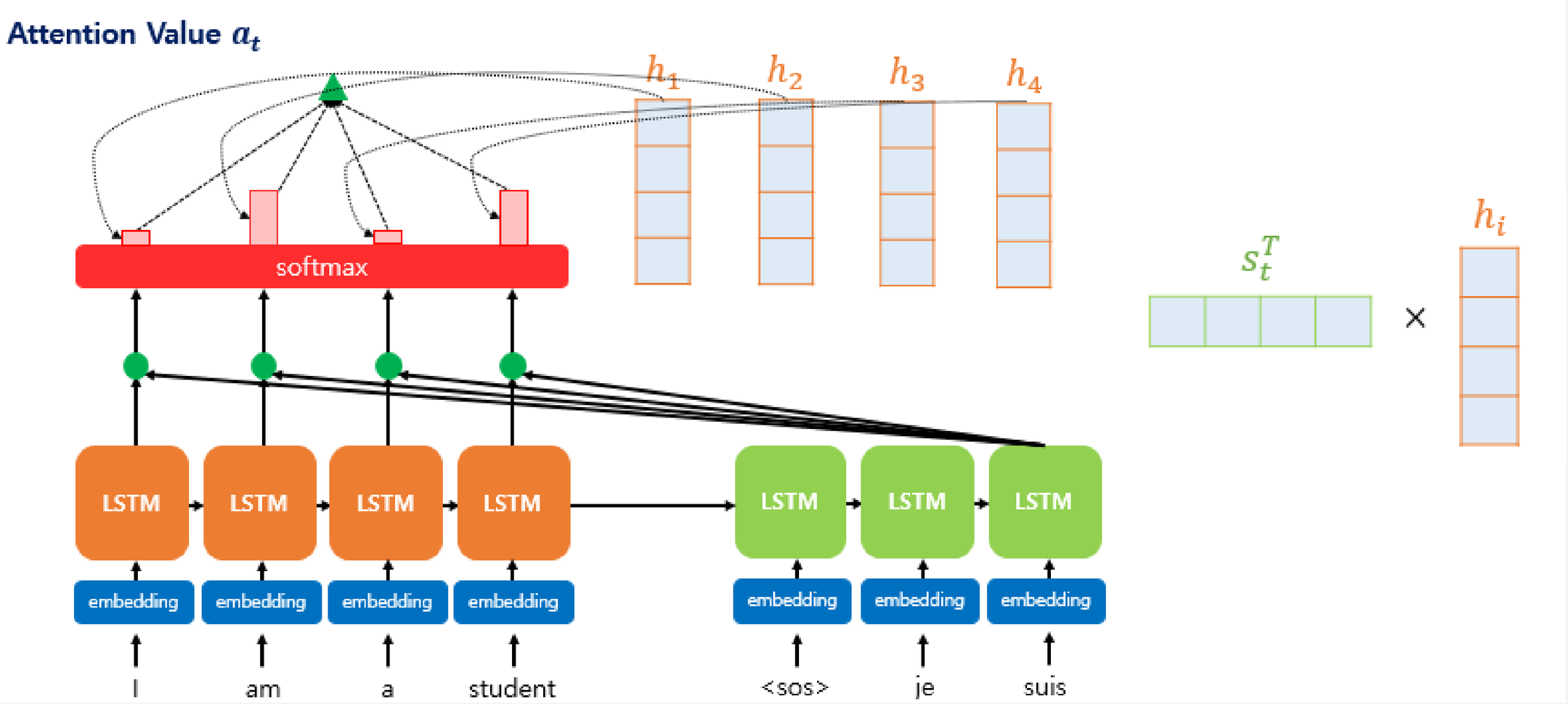
어텐션 메커니즘

Attention mechanism



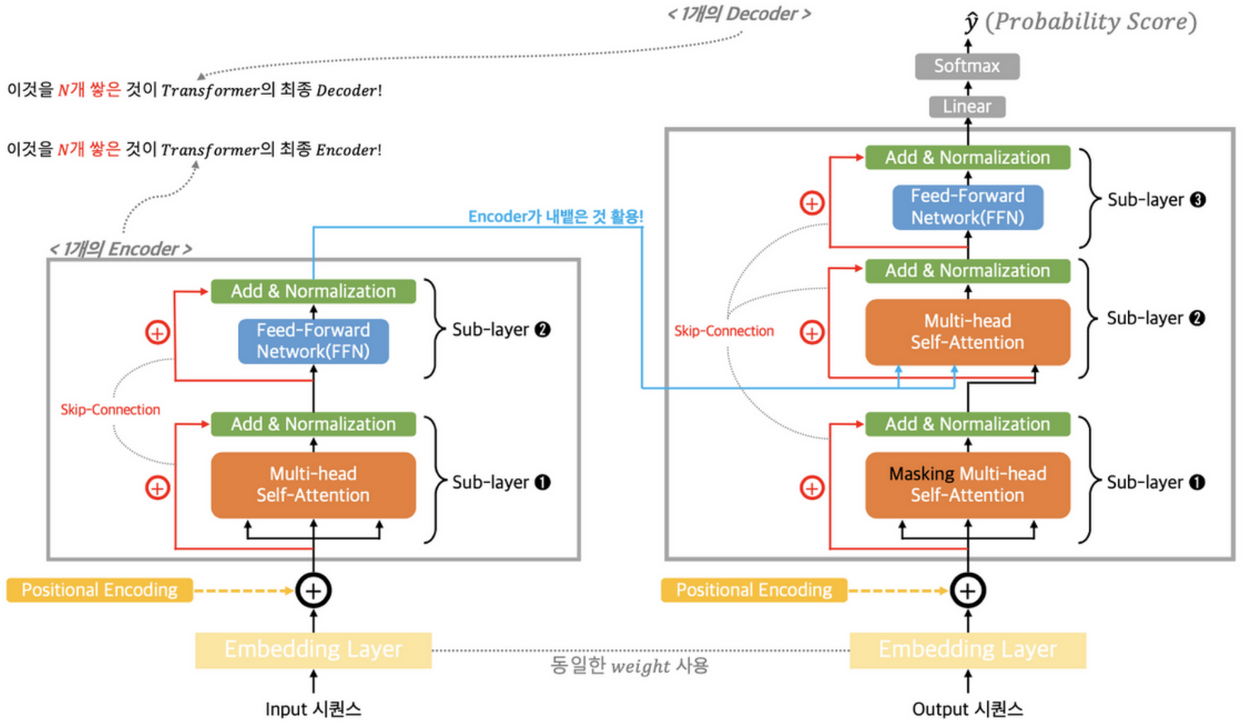
Q에 대해 어떤 K가 유사한지 비교하고, 유사도를 반영하여 V들을 합성한 것이 Attention value이다.

어텐션 메커니즘


Attention Value a_t
h_1
h_2
h_3
h_4
softmax
s_t^T
h_i
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
embedding
embedding
embedding
embedding
embedding
embedding
embedding
I
am
a
student
<sos>
je
suis

트랜스포머

‘Attention is all you need’ (2017) 구글 연구 팀이 발표한 딥러닝 아키텍처이다.
논문 이름에서 알 수 있듯이 기존의 LSTM, GRU 등 RNN 모델에 어텐션 메커니즘을 추가하는 방식과 다르게 어텐션 메커니즘만으로 설계한 모델이다. GPT, BERT 등 대표적인 LLM이 이 구조를 사용한다.



트랜스포머

셀프 어텐션은 Q,K,V가 모두 동일한 어텐션이다.
같은 문장 내 모든 단어 쌍 사이의 의미적, 문법적 관계를 포착해 낼 수 있다.
멀티헤드 어텐션은 최선의 결과를 내기위해 어텐션을 여러 번 시행해 하는 것이다.

Attention(Q, K, V) = Attention Value

Q = Query : 입력 문장의 모든 단어 벡터들

K = Keys : 입력 문장의 모든 단어 벡터들

V = Values : 입력 문장의 모든 단어 벡터들

Scaled Dot-Product Attention

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

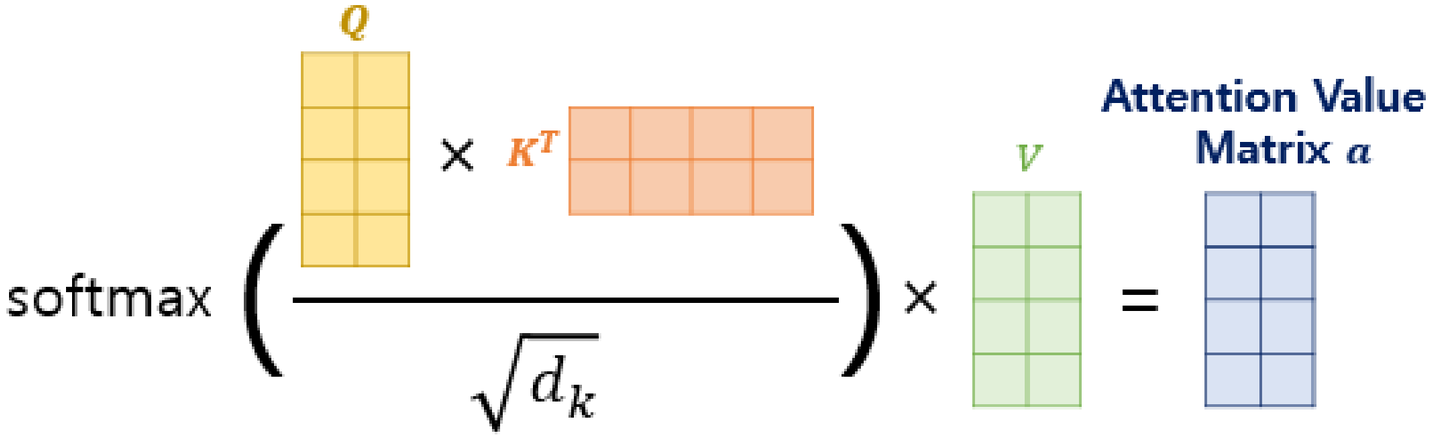
입력 행렬 X
쿼리에 해당하는 가중치 행렬 W-q
키에 해당하는 가중치 행렬 W-k
값에 해당하는 가중치 행렬 W-v

X * W-q = Q
X * W-k = K
X * W-v = V

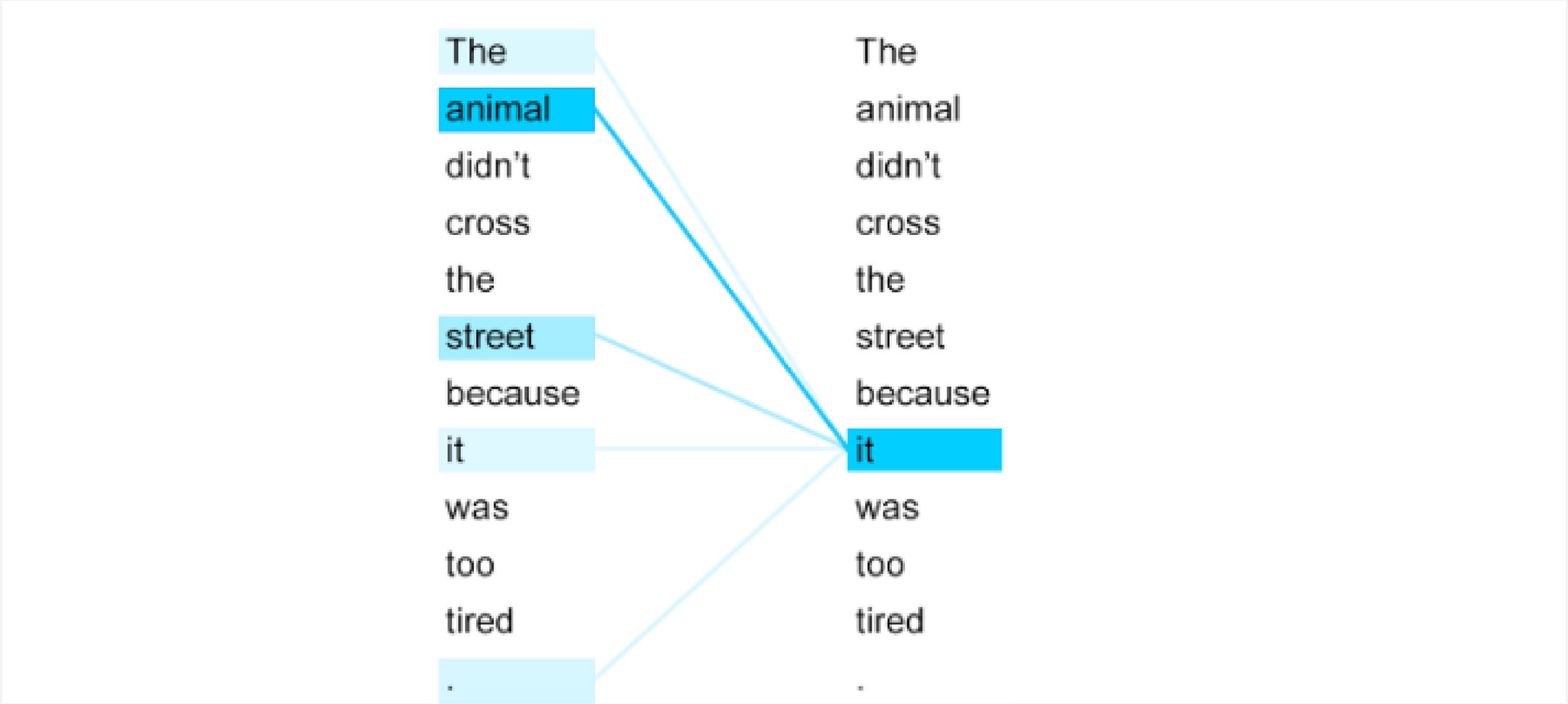
Scaled Dot-Product Attention

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

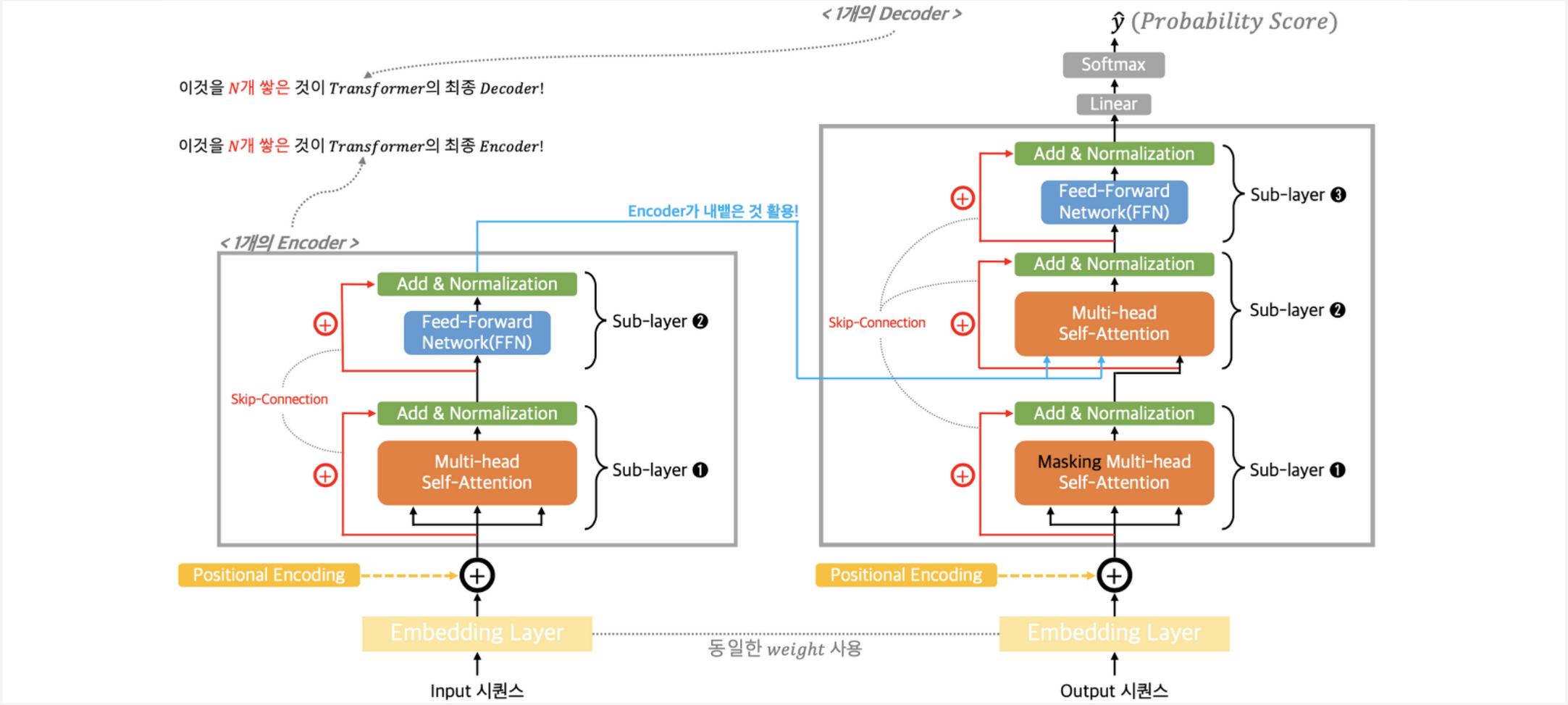
어떤 쿼리와 키가 태스크 수행에 중요한 역할을 하고 있다면, 내적 값이 커지는 방향으로 학습한다.
내적값이 크다는 것은, 두 단위 벡터가 공간 상에 비슷한 위치에 있다는 것을 의미한다.
계산한 내적을 소프트맥스 함수에 통과시켜준다.



Scaled Dot-Product Attention



트랜스포머


< 1개의 Decoder >
ŷ (Probability Score)
Softmax
Linear
이것을 N개 쌓은 것이 Transformer의 최종 Decoder!
이것을 N개 쌓은 것이 Transformer의 최종 Encoder!
Encoder가 내뱉은 것 활용!
< 1개의 Encoder >
Add & Normalization
Feed-Forward Network(FFN)
Sub-layer ❸
Add & Normalization
Multi-head Self-Attention
Sub-layer ❷
Skip-Connection
Add & Normalization
Masking Multi-head Self-Attention
Sub-layer ❶
Positional Encoding
Embedding Layer
동일한 weight 사용
Input 시퀀스
Output 시퀀스

트랜스포머의 장점과 한계

- 그레이디언트 소실을 미연에 방지
- 사용자가 정한 스텝 수까지 학습률을 올렸다가 조금씩 떨어트리는 **웜업** 전략을 사용해 안정적인 학습에 기여함
- 컴퓨터 비전 등 다양한 분야에서의 사용을 확장

- 할루시네이션 발생
- 많은 연산을 필요로 함