Attention Is All You Need

M2023093 한창훈

Attention is all you need

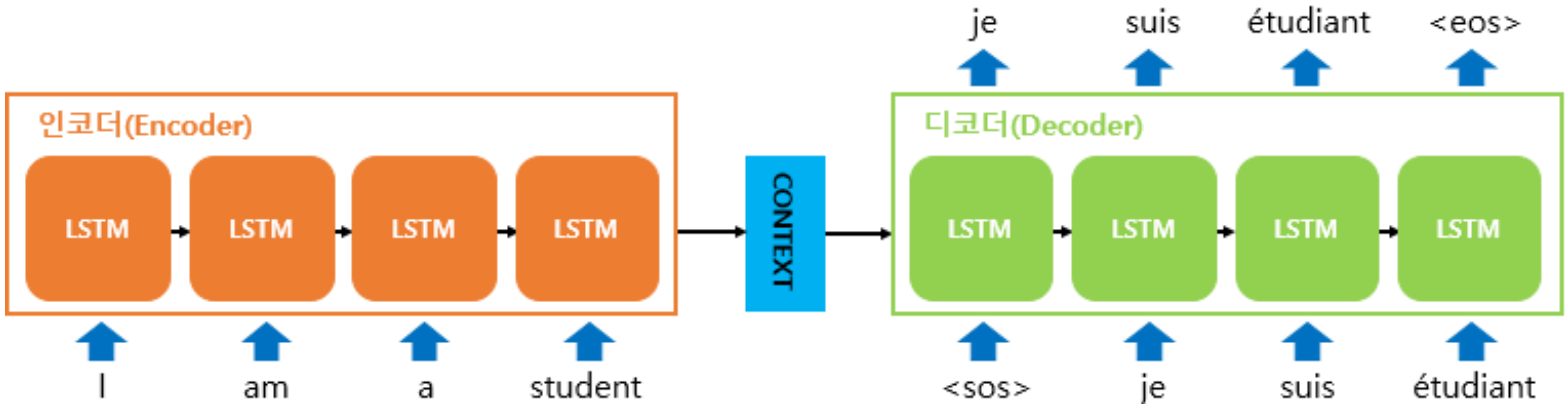
Transformer

- 이 논문에서 번역 Task SOTA 성능을 보이며 등장 (2017년)
- BERT, GPT-3 등에서 사용됨
- 현재는 비전 분야에서도 사용됨

Background

당시 시퀀스 변환 모델: LSTM, GRU등 RNN 기반

- 내부 상태를 사용하기 때문에, 항상 순차적으로 봐야 함
- 병렬화가 불가능, 훈련 속도가 느림
- 이를 해결하기 위한 여러 시도가 있었지만 근본적인 문제 해결 실패



LSTM 기반의 seq2seq 모델

Background

이 문제점을 어떻게 해결해야 할까?

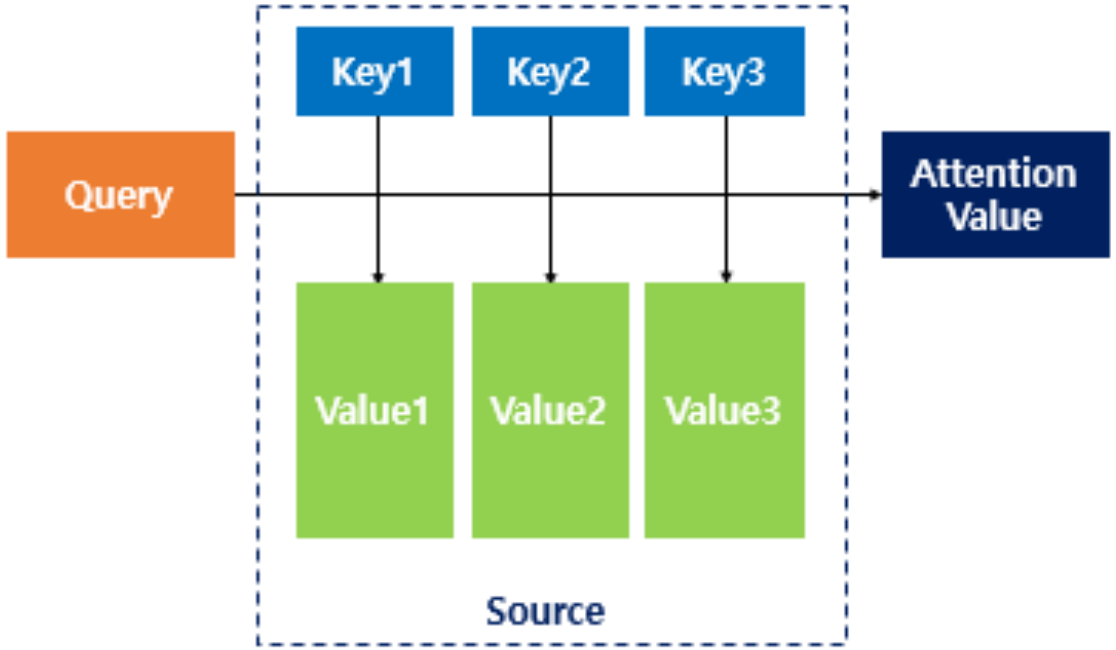
- RNN 기반 seq2seq 모델의 문제점을 보완하는 용도로 Attention 매커니즘이 등장
- 이 논문은 Attention에서 착안해, 모델의 RNN 부분 없이 Attention만을 활용한다는 아이디어
- Attention is all you need!

Model Architecture

Attention

여러 개의 Item에서 각각 다른 양을 참고하기 위한 구조

- Query: 질의 대상 벡터
- Key: 질의 대상과 유사도를 비교할 벡터
- Value: Key와의 유사도에 따라 가져올 값

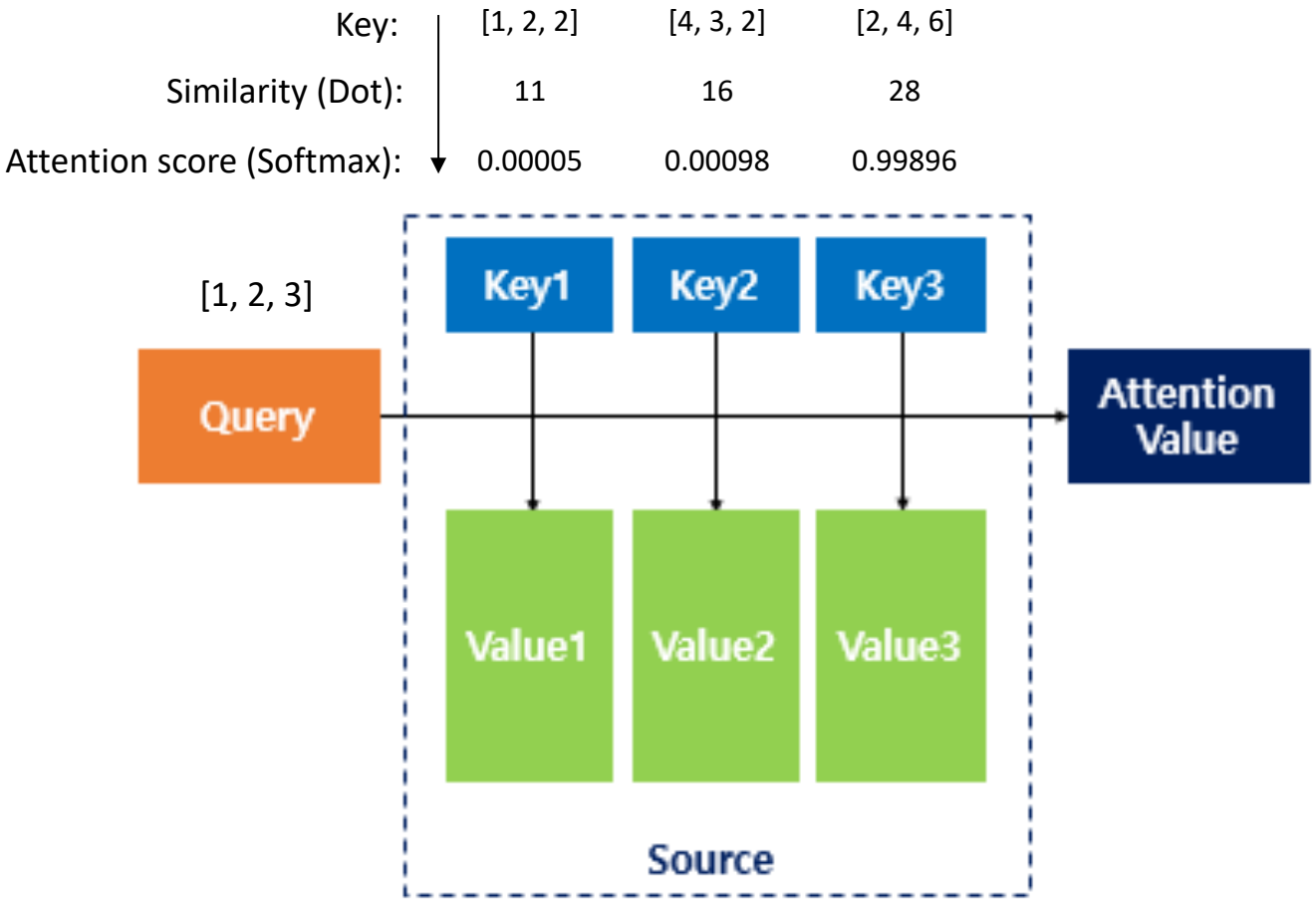


Model Architecture

Attention

- Query: 질의 대상 벡터
- Key: 질의 대상과 유사도를 비교할 벡터
- Value: Key와의 유사도에 따라 가져올 값

- 저렇게 구해진 값을 Attention score라고 함
- Attention score에 비례하게 각 Value들을 참고함



Attention Value = 0.00005*Value1 + 0.00098*Value2 + 0.99896*Value3

Model Architecture

- 문장을 입력
- 토큰화와 Word embedding을 거쳐 Embedding의 리스트로 바꿈
- Embedding의 리스트와 학습시켜야 하는 W와의 곱을 통해

  Q, K, V를 만들어 냄

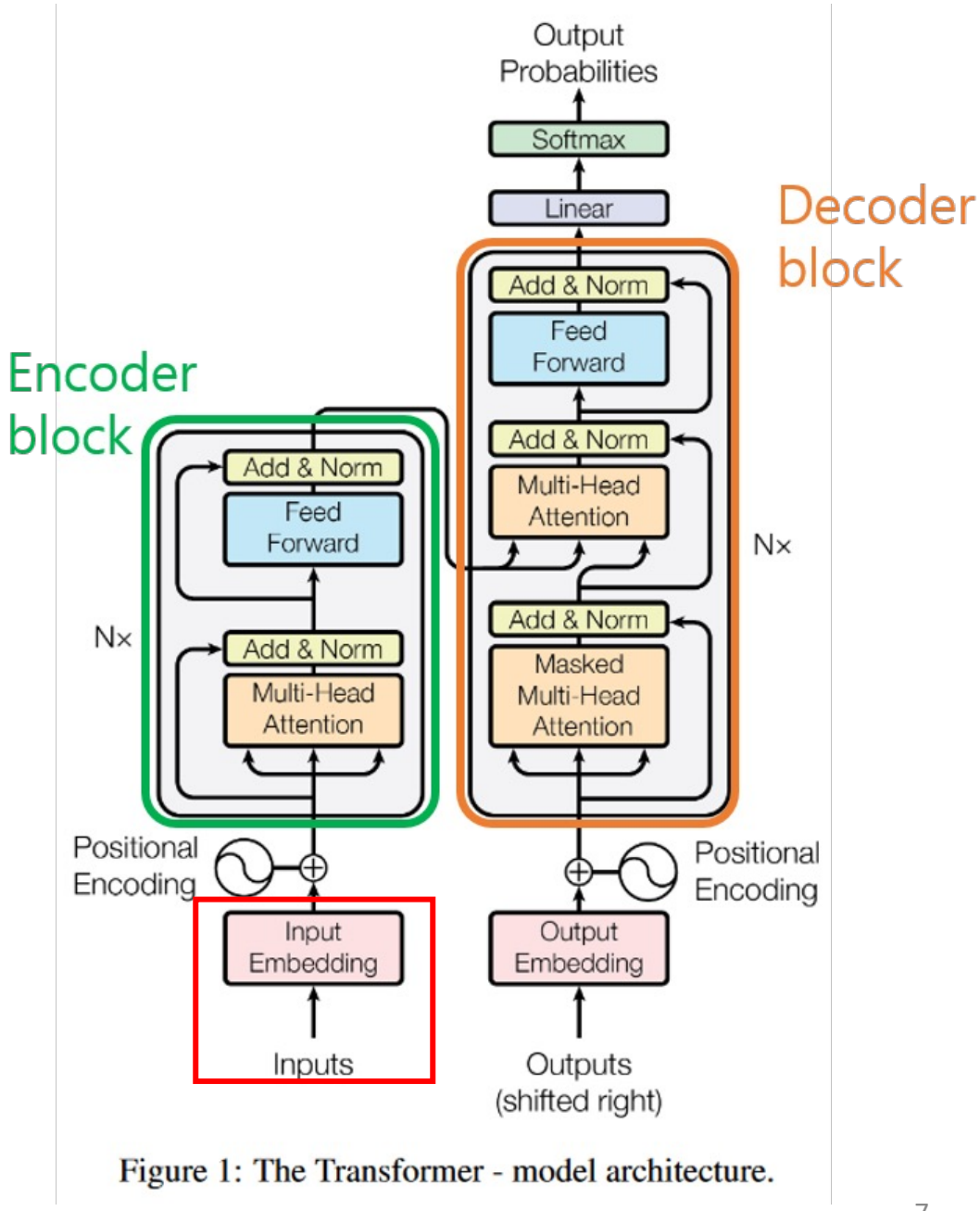
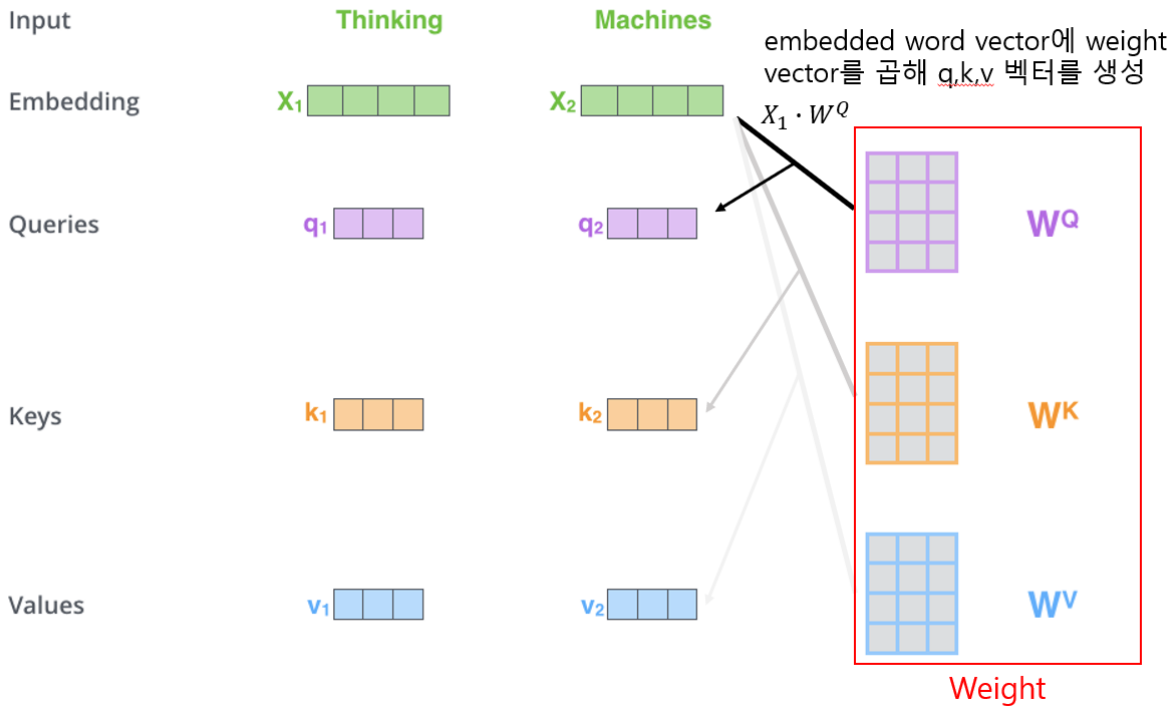


Figure 1: The Transformer - model architecture.

Model Architecture

- 문장을 입력
- 토큰화와 Word embedding을 거쳐 Embedding의 리스트로 바꿈
- Embedding의 리스트와 학습시켜야 하는 W와의 곱을 통해

  Q, K, V를 만들어 냄



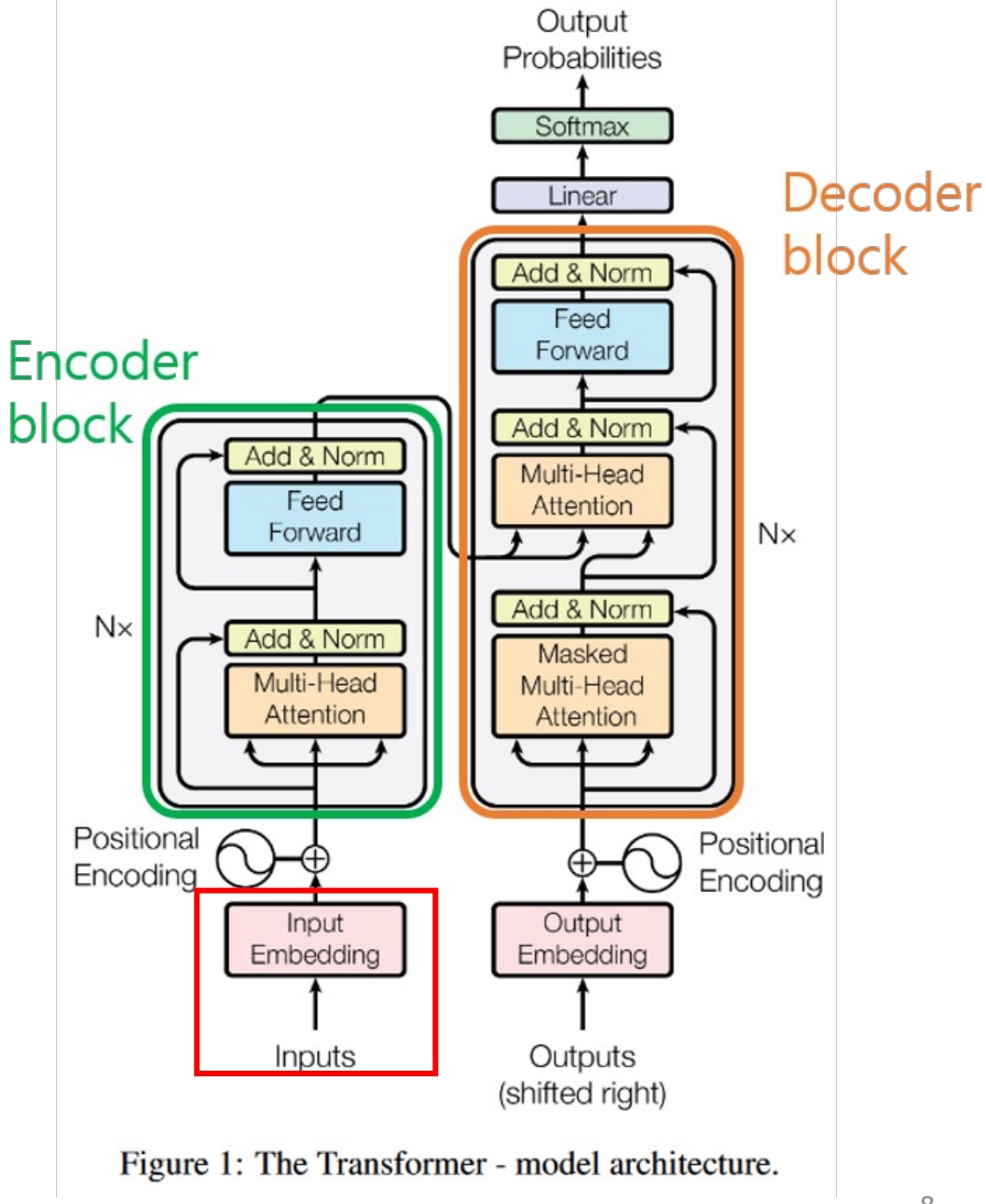


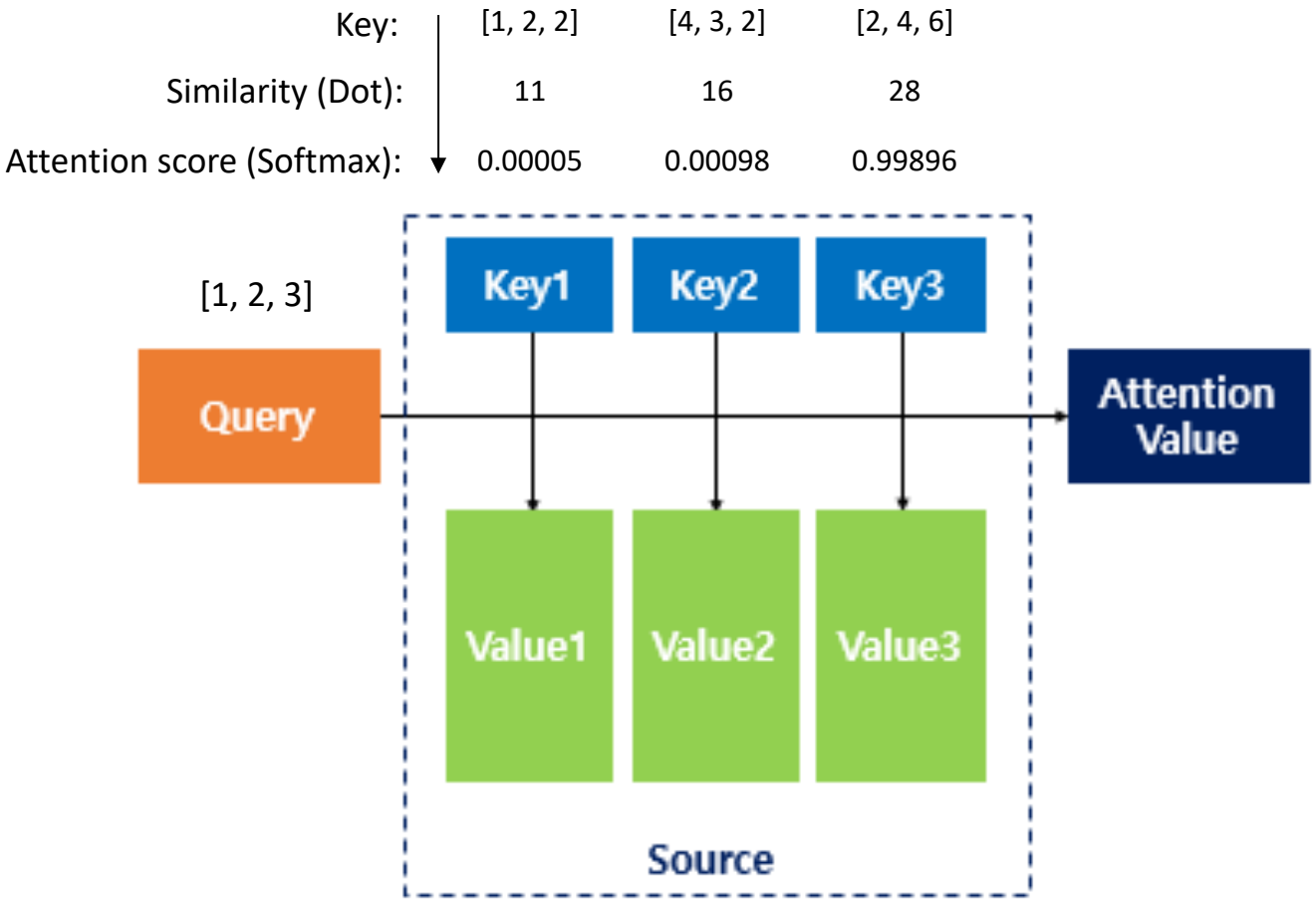
Figure 1: The Transformer - model architecture.

Model Architecture

Self-Attention

- Query: 질의 대상 벡터
- Key: 질의 대상과 유사도를 비교할 벡터
- Value: Key와의 유사도에 따라 가져올 값

- 저렇게 구해진 값을 Attention score라고 함
- Attention score에 비례하게 각 Value들을 참고함



Attention Value = 0.00005*Value1 + 0.00098*Value2 + 0.99896*Value3

이미지 출처

- 3p: https://wikidocs.net/24996