Transformer: 기존의 seq2seq모델의 구조인 인코더-디코더를 따르면서도, rnn 없이 attention만으로 구현됨.

* (RNN은 이전 시각에 계산한 결과를 이용하여 다음 시각의 인풋으로 사용하기 때문에 병렬적인 계산(parallelization)이 힘들고, long range dependency 문제가 존재했음.)

➔ Attention: 인코더가 입력 시퀀스를 하나의 벡터로 압축하는 과정에서 입력 시퀀스의 정보가 일부 손실된다는 단점을 보정하기 위해 사용됨.

Attention을 RNN의 보정을 위한 용도로서 사용하는 것이 아니라 Attention만으로 인코더와 디코더를 만들어보자는 아이디어에서 출발한 Transformer 모델.

왼쪽에 있는 어텐션 함수는 주어진 '쿼리(Query)'에 대해서 모든 '키(Key)'와의 유사도를 각각 구한다. 그리고 구해낸 이 유사도를 가중치로 하여 키와 맵핑되어있는 각각의 '값(Value)'에 반영해주고 유사도가 반영된 '값(Value)'을 모두 weighted sum을 하여 반환한다. 기존 Attention 기법은 입력-출력 간에 대응되는 단어 관계를 파악하는 게 핵심이었다. 하지만 Self-attention 기법은 입력-출력 간이 아닌 입력, 출력 각 시퀀스 내부의 단어들 간의 대응 관계를 파악하는 데 집중한다.

트랜스포머의 기본적인 블락은 두개의 층으로 이루어져있는데, 첫번째는 multi-head attention, 두번째는 Feed-Forward Network 다. Multi-head attention은 attention을 병렬로 수행하여 다른 시각으로 정보들을 수집하겠다는 아이디어다. 여러 개의 헤드들을 연결하는 과정을 거친 다음, 가중치 행렬 w을 곱해주는데, 이로써 생기는 결과물은 처음 인코더에 입력된 문장 행렬과 크기가 같아진다. 트랜스포머는 동일한 구조의 인코더를 쌓은 구조입니다. 논문 기준(Attention is all you need)으로는 인코더가 총6개인데, 인코더에서의 입력의 크기가 출력에서도 동일한 크기로 계속 유지되어야만 다음 인코더에서도 다시 입력이 가능하다.