Attention is all you need

1 배경

1.1 Seq2seq 모델 단점 극복 위해

1.2 RNN 계열 모델이기 때문에 병렬 처리 불가 + 장기의존성 문제

1.3 고정 크기의 벡터에 모든 정보를 압축하기 때문에 정보 손실 문제

1.4 계산 복잡도

2 모델

2.1 인코더 디코더 모두 Positional Encoding, Multi-Head Attention, Feed Forward NN 사용

2.1.1 Scaled Dot-Product Attention

2.1.1.1 Query(Q) – 단어 벡터

2.1.1.2 Key(K) – 문장의 모든 단어들에 대한 벡터 stack한 matrix

➔ QK^T는 한 단어와 모든 단어들의 dop product를 해줌으로써 relation vector 생성

➔ Softmax를 통해 Q가 모든 단어들과 어느 정도의 상관관계가 있는지 표현

➔ Value matrix와 dot product

2.1.2 Positional Encoding = 단어의 위치 정보 나타내기 위함. 위치 벡터 각각의 단어의 반영-> 각 단어의 상대적인 위치 정보 포함-> 시간 정보를 가진 임베딩을 인풋으로 받을 수 있음

2.1.3 Multi-Head Attention = V, K, Q 각각을 h번 다른 linear projection을 통해 변환시키고 병렬적으로 각각의 attention 계산

2.1.4 Feed Forward NN = 문장 내의 정보 추출

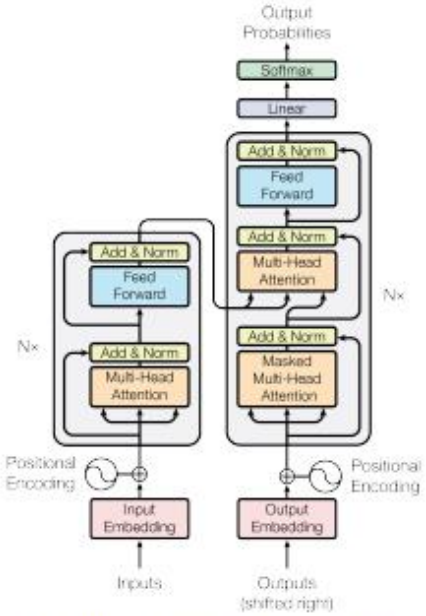

Figure 1: The Transformer - model architecture.

3 장점

3.1 CNN과 RNN보다 빠름

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

3.2 거리에 관계없이 동작 가능

4 실험 결과

4.1 Transformer big model로 SOTA 달성 + low training cost

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	