Week 12

태그	
열	

Attention Is All You Need

Introduction

기존 RNN은 처리해야 할 문장의 길이가 길어지면 성능이 떨어지는 문제가 있고 이를 해결하기 위해 나온 것이 Attention이다. 그러나 기존의 어텐션 메카니즘은 RNN의 접속사 부분에서만 사용됐다. 따라서 해당 논문에서는 어텐션 메카니즘만을 이용하는 모델 구조인 transformer를 제안한다. Transformer는 더 많은 병렬화를 가능하게 한다.

Model Architecture

기본적으로는 encoder-decoder 모델을 따라가는데, self-attention이 쌓인 모습이다. 지점별로 보면 encoder와 decoder 모두 전결합층을 갖는다.

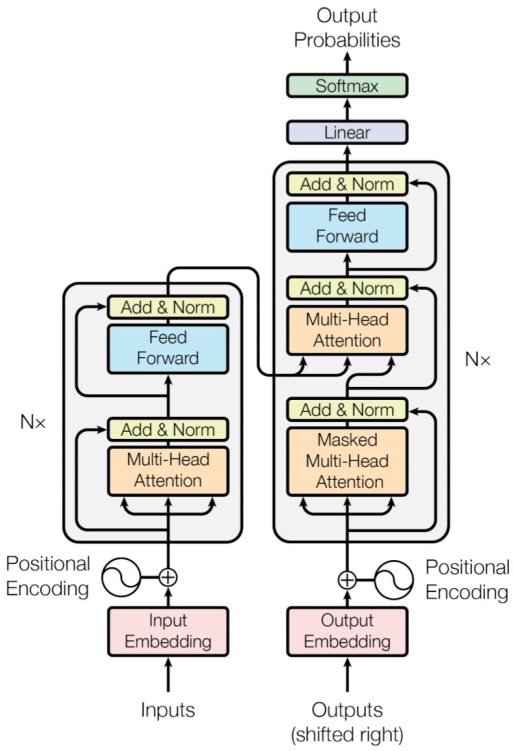

Figure 1: The Transformer - model architecture.

- 인코더: 6개의 동일한 레이어로 구성된다. 각각의 레이어는 두 개의 sub-layer를 갖는다. 첫 번째 sub-layer는 multi-head self-attention 메카니즘이고 두 번째 layer는 간단하게 전결합된 feed-forward 네트워크이다. layer normalization이 일어나므로 모델의 모든 sub-layer와 임베딩 레이어는 512차원이다.
- 디코더: 인코더와 동일하게 6개의 동일한 레이어를 갖는다. 디코더에는 3번째 sub-layer가 있는데, encoder stack의 output에 대해 multi-head attention을 수행한다. 여기서도 layer normalization이 일어난다. position i에 대한 예측은 i보다 이전의 값들에 의해서만 일어나도록 설계되었다.
- Attention: query와 key-value 쌍의 집합을 output과 매핑시키는 것으로 설명할 수 있다. query, key, value는 모두 vector이다. output은 value들의 weighted sum으로 계산되고, 각 value에 할당된 weight는 query와 이에 대응하는 key가 compatibility function을 지나면서 계산된다. 구조를 그림으로 나타내면 아래와 같다.

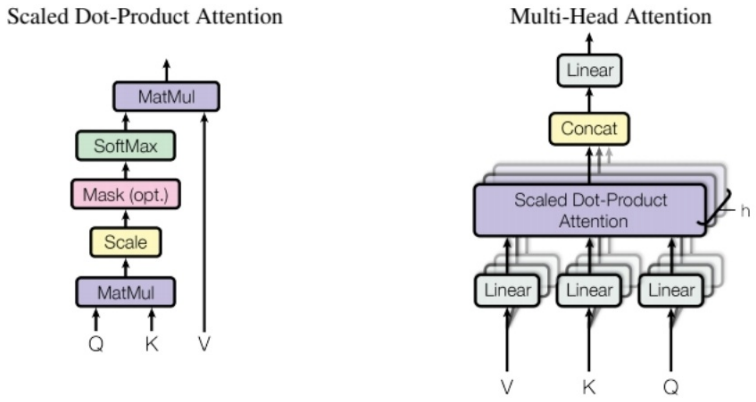

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

- Attention의 적용
  - encoder-decoder attention layer에서 query는 이전의 decoder 층에서 오고 memory key와 value는 encoder output에서 온다. 이는 decoder가 input sequence의 모든 부분을 볼 수 있음을 뜻한다. 가장 기본적인 seq2seq attention 메카니즘을 따라한 부분이다.
  - encoder는 self-attention layer를 가지고 있다. 여기서는 모든 key, value, query가 이전 층의 encoder 결과값으로부터 온다.
  - decoder의 self-attention layer에서 decoder의 각 부분이 해당하는 부분까지의 모든 부분을 커버할 수 있도록 한다.

- Positional encoding: recurrence나 convolution을 사용하지 않으므로 시퀀스의 순서까지 고려해서 학습하도록 하려면 시퀀스에서 토큰의 상대적 혹은 절대적 위치에 대한 정보를 주입해야 한다. 이를 위해 input 임베딩에 positional encoding을 추가한다.

Why Self-Attention

- 다음의 장점을 갖는다.
  - 레이어의 계산 복잡도가 낮다. → 계산이 빠르다.
  - 병렬화 가능성이 높다.
  - long-range dependency를 더 잘 학습한다.