<Attention is all you need paper> 요약

본 논문에서는 Transformer라는 간단한 네트워크 아키텍쳐를 제안한다. CNN과 RNN을 사용하지 않고 Attention에만 의존하는 모델에 대해 설명한다. Sequence가 길어질수록 병렬화가 힘들어진다는 기존의 문제점을 뛰어넘어 Attention은 sequence가 길어져도 decoder가 context를 제대로 참조할 수 있게 되었다. 이 논문의 목표는 순환하지 않는 sequence to sequence에서의 encoder-decoder 모델을 만드는 것이다.

선행연구에서는 LSTM과 GRU 등을 이용한 Recurrent 구조를 사용해왔는데, 병렬화가 어렵고 sequence의 길이가 길어질수록 취약해진다는 한계가 있었다. 그래서 CNN을 활용하려 했으나 인코더와 디코더를 연결하기 위한 추가 연산이 필요하다는 한계가 있었다.

기존의 Attention은 sequence 모델링에서 필수적인 요소로, sequence 내에서 특정 토큰이 다른 토큰과 얼마나 강한 연관성을 가지고 있는지를 embedding으로 표현해주는 것이다. Input과 output sequence에 관계없이 dependency를 학습할 수 있다.

Transformer는 완전히 self-attention에 의존하고 RNN이나 convolution을 사용하지 않는 모델이다. 일괄처리가 되지 않는 Recurrant 구조를 피해 self-attention만으로 입력과 출력의 representation을 계산한다. Input과 output 사이의 global dependency를 학습하고 병렬화에 우수하다.

모델의 전체적인 구조를 보면 encoder-decoder 구조를 가지는데 입력은 인코더를 통해 이루어진다. 인코더는 동일한 6개의 stack으로 구성되어 있다. 각 레이어는 2개의 tj서브 레이어로 구성되어 있다. 첫번째 레이어에서 가장 먼저 self attention을 수행한 후 작업이 끝나면 각 서브 레이어에서 residual connection과 정규화 과정을 거친다. 그리고 Feed forward를 통해 FC를 통과하게 되고, 이것이 N번 반복된다. 이전 레이어로부터 모든 위치를 처리한 정보를 활용한다.

디코더 또한 6개의 동일한 레이어로 구성되는데, 인코더와는 다르게 인코더 출력값에 attention을 수행하는 레이어가 하나 더 있다. 그리고 softmax값을 masking해 후속 position은 attetion을 하지 않도록 조정한다.

Attention에는 크게 두가지 종류가 있다. Scaled dot-product Attention과 Multi-Head Attention이다. Scaled dot-product Attention는 query, key, value 모두 벡터이고, dot product를 수행하는 과정에서 스케일링이 수행된다. Query와 key 사이의 유사도가 높은 value일수록 더 높은 값이 곱해지게 되고, 이것이 가중치로 이용된다. Multi-Head Attention의 경우 각각 따로 계산된 출력들을 concat 해준 후 마지막으로 FC layer에 통과시켜 최종 출력값을 확률값으로 나오도록 한다. 각각 다른 가중치 행렬을 곱해주는 것이 더 좋기 때문이다. Head의 개수는 늘었지만 차원이 줄었기 때문에 단일 head attention과는 계산량이 비슷하다. 이 모델에서는 multi-head attention을 세 가지 다른 방안으로 이용한다.

Transformer가 만들고자 하는 것은 sequence to sequence지만, 위치 정보를 포함해야 한다. 그러므로 Positional Encoding으로 각 단어의 상대적인 위치 정보를 포함해야 한다. 또한, Transformer는 convolution과 recurrence를 사용하지 않고 Multi-Head Attention과 Position-wise Feed-Forward Network를 이용하기 때문에 sequence와 독립적으로 계산된다. 그로 인해 계산을 병렬로 수행할 수 있다. Attention을 활용하면 상대적인 position을 쉽게 학습할 수 있다는 가정하에 sinusoid로 구성한다. Sinusoid로 구성하면 학습과정에서 만나지 못한 긴 문장도 대응이 가능하다. Positional Encoding은 고정 오프셋 K가 있을 때, 선형 변환을 통해 표현 가능해야 한데, Sinusoid로 구성은 선형식으로 표현 가능하다.

Self-Attention을 사용하는 이유는 레이어별로 필요한 연산복잡도가 줄어들기 때문이다 그리고 순차적으로 진행되는 작업의 횟수가 줄어들면서 병렬화가 가능해졌다. 병렬적으로 이루어진 덕에 long-range dependency와 상관없이 빠르게 연산을 수행할 수 있다. 또한, 학습과정에 따라 attention distribution을 확인할 수 있기 때문에, 모델이 설명 가능해진다. 그리고 각 attention이 input sequence에 대한 서로 다른 정보를 저장할 수 있어 서로 다른 task에 특화될 수 있다.

Transformer로 훈련을 시키고 결과를 확인했는데, machine translation, English constituency parsing의 두 task에서 기존 RNN, CNN 기반 모델들보다 더 높은 성능을 보였다.

결론적으로 Multi-Head Self Attention은 convolution이나 recurrent보다 더 빠르게 학습할 수 있고 여러 분야에서 SOTA를 달성했다. Decoder에서 Sequential한 과정을 축소하는 것과 큰 Input과 Output을 어떻게 다룰 것인가가 향후 해결해야 할 과제이다.