모델 심화2: 논문 리뷰
15기 황보진경

Attention Is All You Need

0. Abstraction

sequence 변환은 기본적으로 encoder와 decoder를 포함한 모델로 이뤄져 왔다. 본 논문에서는 Attention Mechanism으로 만들어진 새롭고 간단한 모델인 Transformer를 제안하고자 한다. 이것은 병렬화로 인해 뛰어난 성능을 보여주고 또한, 학습시간도 상당히 줄어들었다.

1. Introduction

기존에 사용되던 LSTM, GRU 등의 RNN 계열의 네트워크들은 히든스테이트와 이전의 히든스테이트, 현재의 input의 조합으로 새로운 시퀀스를 생성한다. 따라서 sequential computation으로 인해 병렬화가 불가능하기 때문에 긴 문장을 처리할 때 큰 어려움이 있었다. 이에 대한 대안으로 attention mechanism이 나왔지만 여전히 RNN과 함께 사용되었는데, 본 논문에서는 recurrence 없이 오직 attention mechanism으로만 이뤄진 모델인 Transformer를 제안하고자 한다. 본 모델은 attention mechanism을 통해 input과 output에 대해 global dependency를 이끌어낸다.

2. Background

sequential computation을 줄이는 목적은 모든 인풋과 아웃풋에 대한 hidden representation들을 병렬로 계산하기 위함이다. Transformer에서는 attention-weighted position의 평균에 의한 비용이 있을지라도, Multi-Head Attention의 효과로 연산의 수를 줄일 수 있다. Self-attention은 sequence의 표현을 계산하기 위해 한 문장에서 다른 포지션들간의 attention-mechanism을 의미한다. Transformer는 RNN이나 CNN 사용 없이 전적으로 self-attention을 사용한 첫번째 모델이다.

3. Model Architecture

대부분의 neural sequence transduction model은 encode-decoder 구조를 가진다. 이것은 input sequence → continuous representation z → output sequence 형태이며, 각각의 스텝에서 이전에 생성한 값을 추가적인 인풋으로 사용한다. Transformer는 인코더와 디코더 모두 stacked self-attention, point-wise, FC layer로 이뤄져 있다.

인코더는 6개의 동일한 레이어로 구성되어 있다. 각각의 레이어는 milti-head self-attention과 positionwise Fully connected feed-forward network로 이뤄져 있다. 디코더도 6개의 동일한 레이어로 이뤄져 있으며, 인코더와 동일한 구조에다가 masked multi-head attention이 추가되어 있다. masked multi-head attention은 뒤에 나타나는 것들에 대해 attention을 하지 않도록 masking을 추가한 것이다. 인코더와 동일하게 각각의 서브레이어마다 layer normalization 뒤에 residual connection을 추가하였다.

Attention은 query와 key-value 쌍을 output으로 매핑하는 것이다. output은 value들의 가중합으로 표현된다. 이때의 가중치는 키에 일치하는 query의 compatibility function에 의해 계산된다. 아래와 같이 나타낼 수 있다. 본 논문에서는 Scaled Dot-Product Attention을 사용하는데, 그 식은 아래와 같다.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

즉, 모든 키에 대해서 쿼리와 내적을 한 후, root(key의 차원)으로 나눠준다. 그리고 sofrmax를 취한 후, value를 가중합한다. dot-product attention은 scaling factor만 빼면 위 함수와 동일하다. Additive attenttion은

히든레이어로 이뤄진 feed-forward network를 사용하여 계산한다. 두개의 이론적 복잡도는 비슷하지만 dot-product attention은 시공간 측면에서 더 좋다.

Multi-head attention은 다른 위치에서 다른 표현공간으로부터 attention을 수행한다. 쉽게 말해, h번 attention을 수행하고, 그 결과를 concatenate한다. 각각의 head들의 차원이 줄어들기 때문에 전체 계산비용은 full dimension의 single attention과 비슷하다.

Transformer에서 multi-head attention은 세가지 방식으로 사용된다. ① encoder-decoder layer에서 쿼리는 이전 디코더 레이어로부터 오고, memory keys와 value들은 인코더의 output으로부터 온다. 즉, decoder는 인풋 시퀀스에 있는 모든 위치에서 attention이 가능하다. ② 인코더의 self-attention layer의 같은 장소에서 key, value, query가 오기 때문에 인코더의 각각의 위치는 이전 레이어의 인코더의 모든 위치를 attention할 수 있다. ③ 디코더의 self-attention은 마찬가지로 모든 위치에 접근할 쑤 있기 때문에 making을 통해 잘못된 connecction을 막았다.

인코더와 디코더는 다른 포지션에서의 선형변환과 ReLU activation을 포함하는 feed-forward network를 포함한다. 또한 Positional Encoding을 통해 같은 단어이더라도 위치에 따라 다른 벡터로 변환된다.

4. Why Self-Attention

Self-attention을 통해 세가지 원하는 바를 이룰 수 있었다. ① 레이어별 총 연산 비용 감소 ② 병렬화 가능한 연산 ③ 네트워크에서 long-range dependencies. long-range dependency에 영향을 미치는 요인은 forward와 backward signal의 길이이다. 이 길이가 짧을 수록 좋다. 그러므로 다른 레이어 타입으로 이뤄진 네트워크에서 임의의 input과 output의 최대 길이를 비교하였다. 그 결과, self-attention layer 상수시간안에 연산한 반면, recurrent layer는 O(n)의 연산을 필요로 했다. 또한, 부가효과로, self-attention은 더 해석가능한 모델을 산출한다.

6. Results

English-to-German Machine Translation에서 big transformer model은 기존의 best model보다 2.0BLEU 높은 성능을 기록하였다. 학습은 8개의 P100 GPU로 3.5일 걸렸다. English-to-French translation task에서는 기존의 모델보다 높은 성능르 보이면서 학습비용도 1/4수준으로 적게 소요되었다.

7. Conclusion

본 연구에서는 multi-head self-attention으로 recurrent layer를 대체한 최초의 sequence transduction model이다. 번역에서, Trnasformer는 기존의 모델들(앙상블 포함)보다 상당히 빨리 학습되며 새로운 SOTA를 이뤄냈다. attention-based model을 적용한 미래의 연구들을 기대한다.