Attention Is All You Need

김민규

Abstract

- 성능 좋은 변환(번역) 모델은 인코더와 디코더를 포함한 복잡한 recurrent 또는 convolutional 신경망에 기반을 두고 있다. 최고 성능을 내는 모델 역시 attention mechanism을 사용하여 인코더와 디코더를 연결한다.
-

Introduction

- RNN, LSTM, GPU 등은 sequence 모델링과 언어모델 등 변환 문제, 기계번역 등의 문제에서 뛰어난 성과를 보였다.
- Recurrent 모델은 보통 입력과 출력의 symbol position에 따라 계산을 수행한다. 계산 단계에서 위치를 적절히 맞추기 위해 이전 상태와 위치 t의 함수인 은닉상태를 생성한다.
- Attention mechanism은 입력과 출력 sequence의 거리에 상관없이 의존성을 모델링함으로써 다양한 과제에서의 sequence 모델링과 변환 모델에서 매우 중요한 부분이 되었다. 그러나 거의 대부분의 경우 recurrent 네트워크와 함께 사용되고 있다.

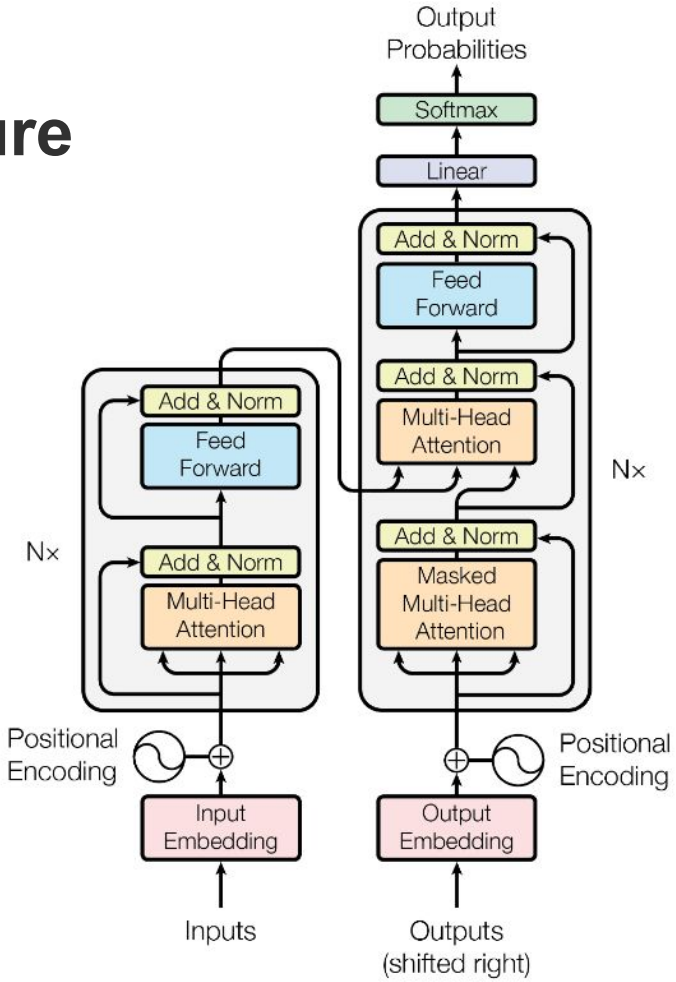
Background

- 연속적 계산을 줄이려는 노력은 Extended Neural GPU, ByteNet, ConvS2S 등의 모델을 탄생시켰으나 이들은 전부 CNN을 기본 블록으로 사용한다. 이러한 모델들은 임의의 위치의 input-output 사이의 관련성을 파악하기 위해서는 거리에 따라(선형 또는 로그 비례) 계산량이 증가하며, 이는 장거리 의존성을 학습하기 어렵게 한다.
- Transformer는, 이를 상수 시간의 계산만으로 가능하게 하였다.
- intra-attention으로도 불리는 Self-attention은 sequence의 representation을 계산하기 위한 단일 sequence의 다른 위치를 연관시키는 attention mechanism이다. Self-attention은 많은 과제들에서 사용되었으며 성공적이었다.

Model Architecture

- Transformer는 크게 인코더와 디코더로 나뉘며, 인코더는 입력인 symbol representations (X1,...Xn)을 continuous representations z = (z1, …, Zn) 으로 매핑한다.
- Z가 주어지면, 디코더는 한번에 한 원소씩 sequence (Y1, … , Yn)을 생성한다.
- 각 단계는 자동회귀(auto-regressive)이며, 다음 단계의 symbol을 생성할 때 이전 단계에서 생성된 symbol을 추가 입력으로 받는다.
- Transformer은 인코더와 디코더 모두에서 쌓은 self-attention과 point-wise FC layer을 사용한다.

Model Architecture



Encoder and Decoder Stacks

- 인코더는 N = 6개의 동일한 레이어로 구성되며, 각 레이어는 아래 두 개의 sub-layer로 이루어져 있다.
- muti-head self attention mechanism
- simple, position-wise fully connected feed-forward network
- 각 sub-layer의 출력값은 LayerNorm(x) + Sublayer(x)이고, Sublayer(x)는 Sub-layer 자체로 구현되는 함수이다.
- 이 residual connection을 용이하게 하기 위해, embedding layer를 포함한 모델의 모든 sub-layer는 d = 512차원의 출력값을 가진다.

Attention

- Attention 함수는 query + key-value -> output으로의 변환을 수행한다.
- query, key, value, output은 모두 벡터이다.
- output은 value들의 가중합으로 계산되며, 그 가중치는 query와 연관된 key의 호환성 함수에 의해 계산된다.