Attention Is All You Need.

Abstract

1. Introduction

2. Background

3. Model Architecture

    3.1 Encoder & Decoder Stacks

    3.2 Attention

        3.2.1 Scaled Dot-Product Attention

        3.2.2 Multi-Head Attention

        3.2.3 Applications of Attention in our Model

    3.3 Position-wise Feed-Forward Networks

    3.4 Embeddings and Softmax

    3.5 Positional Encoding

4. Why Self-Attention

5. Training

    5.1 Training Data and Batching

    5.2 Hardware and Schedule

    5.3 Optimizer

    5.4 Regularization

6 Results

    6.1 Machine Translation

    6.2 Model Variations

    6.3 English Contituency Parsing

7. Conclusion

Abstract

- Simple network architecture … Transformer
- based solely on attention mechanisms!

- Experiments (machine translation):
  - superior in quality
  - more parallelizable
  - require significantly less time to train

- WMT 2014, best BLEU

- generalizes well!
  - [*1] Constituency parsing

1. Introduction

Recurrent models

- typically factor Computation along the symbol position of the input and output sequences.
- Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as
  - a function of previous hidden state h_{t-1}
  - input for position t.

↓

preclude parallelization within training examples, critical at longer sequence lengths!

(improvements : factorization tricks & conditional computation)

- Attention mechanism

  : allow modeling of dependencies without regard to their distance in the input or output sentences.

Transformer

- rely entirely on an attention mechanism to draw global dependencies between input & output
- more parallelization
- SOTA in translation quality
- 12 hrs training with 8 P100 GPUs.

2. Background

- to reduce sequential computation ; CNN
  - Extented Neural GPU
  - ByteNet
  - ConvS2S

the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions.

- linearly for ConvS2S
- logarithmically for ByteNet

★ Transformer ★

- reduced to a constant number of operations,
- reduced effective resolution due to averaging attention-weighted positions

↓

- counteract with Multi-Head Attention
- Self - Attention

*1. Constituency parsing

: 문장이 구 단위로 묶여가면서 구조를 이뤄가는 방법

어순이 고정적인 영어에서 쓰임.