Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Attention Is All You Need

1. Paper Summary

In sequence modeling and transduction problems, recurrent neural networks and its variants that embody the attention mechanisms are regarded to be the state-of-the-art approaches. These typically inherent sequential nature so that it precludes parallelization within training examples, which is critical at long sequences due to the memory constraints. There were some previous works to reduce the sequential computations through conventional neural networks as basic building block computing hidden representations in parallel for all input and output positions. However, it turns out to be computationally inefficient in the sense that the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions.

Therefore, as a breakthrough, the author proposed a new simple network architecture that is named as 'Transformer' which relies solely on attention mechanism without using recurrence and convolutions. Eventually, the experiment shows that it outperforms the previous methods in English-to-German and English-to-French translation task with little training time significantly utilizing more parallelization.

The transformer follows an encoder-decoder structure using stacked multi-head attentions and point-wise fully connected layers for both the encoder and decoder. Here, the multi-head attention which applies scaled dot product attention, allows the model to jointly attend to information that is project to different representation subspaces at different positions. In particular, there is an additional sub-layer in decoder, which performs multi-head attention over the output of the encoder stack. As a result, by utilizing the attention mechanisms in 3 different ways (Encoder, Decoder, Encoder-Decoder), each position in the encoder can attend to all positions in the previous encoder layer of the encoder and each position in the decoder can not only attend to all positions in the previous decoder layer, but also to all positions in the input sequence.

Below are the several additional side remarks.

- A residual connection is employed around each of these layers that is followed by the layer normalization.

- The masking is used in the first multi-head attention in decoder to prevent leftward information flow to preserve the auto-regressive property so that the predictions for position i can depend only on the known outputs at positions less than i.

- Fully connected feed-forward network consists of two linear transformation which are the same across different positions, but use the different parameters from layer to layer.

- Learned embeddings/pre-softmax linear transformation are used to convert the input and output tokens to vectors and the decoder output to predicted next-token probabilities and their weight matrices are shared.

- Positional encoding is used at the bottom of the encoder and decoder stacks by injecting some information about the relative or absolute position of the tokens in the sequence to make the model to use the order of the sequence.

2. Discussion

Here I would like to offer 2 discussion points. To begin with, when applying multi-head attention, how can the better representation can be learned rather than simply applying linear projection? This can be resolved through autoencoder where various information theoretic metrics are provided. I suggest the disentanglement among the features dimension to be encouraged to further improve the translation quality, which makes it more independent and interpretable. Next, layer normalization is performed before the multi-head attention? I guess this would make the training process more stable.