Attention Is All You Need

December 15, 2023

1 Summary

0. Existing sequence transduction models use recurrent or convolutional networks while the best performing ones also include an attention mechanism. This paper proposes a simple algorithm -Transformer - using only an attention mechanism and it outperforms all other existing solutions in language translation. This also comes with reduced training costs as it is parallelizable.

1. Recurrent models' sequential computation nature precludes parallelization. Attention models were developed to deal with this but existing literature still used recurrent network with such models.

2. Although convolutional network approach addresses parallelization, operations required in relating two input or output signals grows with distance. In the Transformer this is constant but with reduced resolution an effect which is countered by Multi-headed Attention.

3. **Model Architecture.** The input is passed through encoder and then the decoder each of which has 6 layers. Each layer in encoder has two sub-layers first being the multi-head attention layer and then the position encoded fully connected feed forward network. A layer norm is also employed.

   The decoder layers are similar to the encoder layers except there is an added multi-head sub-layer to process output from encoder layer. Also one layer is masked so that predictions for a position can only depend on known outputs before that position.

   Attention can be seen as a mapping from query and set of key-value pairs to an output. In matrix form the scaled dot product attention is given by

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

   Multi-head attention involves linear projection of query, key, value pairs among multiple heads and computing parallelly. These are then concatenated to form output. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.

In this attention model, the decoder can to attend every position from the input.

Positional Encoding. Since there is no use of recurrence or convolution, positional information is not used. Hence, positional encoding is added to the model to learn from it.

4. Why Self-Attention. It is motivated by computational complexity, computation parallelization and more importantly path length between long-range dependencies in the network. As a side benefit this network also yields interpretable heads in the language translations task presented in the paper.

5. **Training and Results.** The model is trained with German-English translation and obtained best results to date and cost effectively.

2 Analysis

This paper picks up the ‘attention’ idea that was being used to predominantly improve recurrent and convultion approaches and makes it the centre stage of this network. The design idea is well motivated in dealing with long-range dependencies computation parallelization. However, it yielded surprisingly good results that too with less costly training. However it is unclear if any design principle in this paper can be learnt to apply easily to some future problem.