Jacob Devasier

4/23/2021

Review of Attention is All You Need

In this paper, the authors approach the problem of long-term dependencies in RNNs. As is already known, RNNs suffer from not being able to retain information from the beginning of a sequence. Because of this, LSTMs were created to help with filtering out important and unimportant information in the hidden state; however, even LSTMs, by nature of the recurrency in the architecture, suffer from the same problem.

In encoder-decoder networks, the model must traverse through the entire input sequence before generating an output sequence, thus leading to the aforementioned problem. To solve this problem, previous work has introduced attention, allowing the model to essentially look back at the earlier input states rather than getting only a final input state. While this does solve the problem of long-term dependencies, it retains the problem of sequentially processing each token, thus lacking the ability to run in parallel. The transformer, which uses an attention mechanism in addition to a positionally encoded feedforward network, solves all of these problems.

Because the transformer uses a feedforward layer instead of a recurrent layer, it is able to process all of the input sequence at once instead of sequentially. This allows the model to be run and thus trained in parallel allowing for a faster training time. In addition, because the RNN layers have a time complexity proportional to the square of the dimensionality of the input, RNNs are limited in the amount of information that can be represented in each token. On the

other hand, the transformer has a time complexity proportional to the square of the sequence length, allowing for the dimensionality of the input to be increased drastically. As we have seen with GPT-3, which uses the transformer, using roughly 12,000-dimension representations of each token.

The transformer paper is revolutionary and has led to drastic increase in performance for nearly every NLP task, but it does have some flaws. In particular, the limitation on sequence length prevents the transformer from being applied to large texts. In addition, a sequential understanding of the input can still be useful, e.g., when predicting the output of some piece of code. In this example, the transformer would lack the ability to reason through the code sequentially.