Paper: *Attention Is All You Need*

Razwan Ahmed Tanvir

Date: 03/04/2022

Quote *"... An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. "*

Overview Sequence to sequence models based on recurrent neural networks have encoder and decoder in them. Also, some of the best models use attention mechanism to improve sequence to sequence prediction. In this paper, the authors proposed an architecture which they named the "Transformer". This model is solely built on utilizing the attention mechanism. They did not use the recurrence and convolutions in their model. They claimed their model is superior in terms of translation quality and also it is parallelizable. In general recurrent models compute along the positions of the input and output sequence. Moreover, the attention mechanism is currently an important part of the sequence modeling which even works with longer input sequence. In this paper, the author used Encoder an Decoder stacks where encoder and decoder are of 6 layers. They also modified the self-attention sub layer. This masking makes sure that predictions of i^{th} position outputs at positions which is smaller than i. Now, attention is a mapping which maps a query and a key-value set to an output. Output of the attention is a weighted sum which assigns the amount of attention for certain position in the sequence. The authors discussed about two types of attention- 1) scaled Dot-product Attention, 2) Multi-Head attention. They also used fully connected Position-wise Feed-Forward Networks. They used softmax to generate possibilities of the next token. The authors used three types of regularization.

This paper has major accomplishment in terms of novelty and accuracy of the proposed model. The authors used the WMT 2014 English to German dataset to measure the performance of their model. They reported a state-of-the-art BLEU score of 28.4. The model took 3.5 days to train. Moreover, the transformer model reduced the training cost compared to recent models.

Intellectual Merit Transformer model for sequence to sequence prediction provided state-of-the-art output in terms of performance and translation quality. The authors proposed a novel transformer architecture which used the attention mechanism to predict sequence. This research is organized and the authors provided reasoning for the performance gained by the model. Moreover, this research is from google and they have access to required hardware to train such huge model.

Broader Impact This research improved the overall output of the translation tasks and provided best results with this novel Transformer model. The authors used WMT 2014 English to German dataset which can be found online, and also the code to train this model is given within a github repository. There are no under represented group among the authors.

Keywords Natural Language Processing, Attention, Transformers, BLEU

Discussion Questions

- The authors used the WMT 2014 dataset which has English to German translation, but I wonder if the performance would vary if we change the language pairs.
- The authors did not discuss about why they did not use alignments and if it really help improve the performance of the model.

Table 1: Grade deductions by section

Overview	Intellectual M.	B. Impact	Keywords	Questions	Is Online?