Paper: *Transformer: A Novel Neural Network Architecture for Language Understanding*

Razwan Ahmed Tanvir

Date: 03/14/2022

Quote *"... we show that the Transformer outperforms both recurrent and convolutional models on academic English to German and English to French translation benchmarks."*

Overview Neural Networks approach, specifically Recurrent Neural Networks (RNN) are proven to be effective for the language processing tasks such as machine translation, question answering and language modeling. However, the authors, in previous paper named "Attention Is All You Need", proposed a transformer model which outperformed the convolutional and recurrent neural network approaches for natural language processing in English-German and English-French translation benchmarks. Moreover, the proposed Transformer model was also improved multiplicably in terms of hardware optimization and speed up during training. In comparison to previously developed models, proposed model showed improved BLEU scores. Neural Networks works on a fixed length sentences and aggregate the information of the input tokens. On the other hand, Recurrent Neural Networks process the input sequence from left to right or even right to left manner and make decisions in multiple steps. However, the Transformer model performs constant amount of steps to take decision. In addition this model uses self attention mechanism which builds relationship among all the input tokens. This ensures that the whole context of the given sequence is retrieved. By comparing each word with every other word, an attention score is generated. In this way, the model gains the context of the input sequence.

Apart from the better translation quality and performance boost during training, the other intriguing outcome of this research was that, the authors could show what part of the sentence is getting more attention. Transformer models, hence, provide better results in modelling languages and also machine translation as it can better accumulate the context of each word in relation to the other words in the input sentence.

Intellectual Merit Improved performance and hardware optimization were offered by the transformer model. Furthermore, this model produced state-of-the-art results in sequence to sequence prediction tasks. This model incorporated the attention mechanism which drastically performed better in terms of establishing meaningful relation among the words in the input sequence. The blog author Jakob Uszkoreit is from the Google Brain team and google has built a great team to carry this research with the plentiful resources necessary for this research.

Broader Impact Transformer model clearly performed better than any other recent approaches to tackle language modelling and translation tasks. This research is always referred in the field of natural language processing research. The authors Ashish Vaswani, Noam Shazeer and Łukasz Kaiser are from google brain team and there present an academic Aidan N. Gomez, from the University of Toronto who worked with the Google brain team. Moreover, the codes are available to use for public and the dataset is also available on the internet.

Keywords Natural Language Processing, Attention, Transformers, RNN, CNN

Discussion Questions

- The authors did not mention about the performance in other datasets for their model. It should be analyzed that how the model works in different language pair.
- The author could discuss about the custom attention and the ways to identify if custom attention could improve the models performance.

Table 1: Grade deductions by section

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