Rutuja Durge

21070521062

Sem – 7 A

ATTENTION IS ALL YOU NEED – SUMMARY

The document is written by Ashish Vaswani, Noam Shazeer, Niki Parmer, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhim. There are various instances where we do not understand the language of the opposite party, be it a foreign person, a distant friend or a computer. So, to perform effective communication we need a translator or a mediator. This paper focuses on building a Transformer Model using different architectures to obtain the best performance that performs the same function.

The model aims to do something better than the conventional neural networks like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), the RNNs are designed to handle sequential data by maintaining a hidden state that captures information about previous elements in the sequence. Now this sequential processing can be slow and computationally expensive. If the user only needs translation for only a particular part of the process, then the translation if performed on the entire process will be wasted and unnecessary. Hence, the concept of attention was used in this model. Attention mechanisms allow models to focus on specific parts of the input sequence when generating each part of the input sentence which are relevant. For the transduction of tasks like reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations, self-attention is used (also known as intra-attention) which an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Usually, processes use sequential processing for the execution of tasks, but this paper uses multi-head self-attention and point-wise, fully connected layers for both the encoder and decoder. Which allows for significant parallelization, which means processing the tasks parallel to one another that in turn improves training efficiency. Further, the paper explains the two models of Attention, i.e Scaled Dot-Product Attention and Multi-Head Attention. But the Transformer Model uses the Multi-Head Attention Model.

The results of this model were amazing, the Transformer achieves state-of-the-art results on the WMT 2014 English to German and English-to-French translation tasks, which are benchmark translation tasks from the Workshop on Machine Translation (WMT) which is a prestigious annual competition in the field of machine translation. The Transformer model also had high BLEU scores and less training time as compared to previous models, BLEU score (Bilingual Evaluation Understudy) is a metric for evaluating the quality of text that has been machine-translated from one language to another. Lastly, one of the applications of this Transformer is the English constituency parsing, wherein the output is subject to strong structural constraints and is longer than the input. The results showed that despite the lack of task-specific tuning, the model performs well by giving better results than all previous models.