Attention Is All You Need [Summary]

In the paper titled, “Attention Is All You Need” by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin the authors introduce the new model called the Transformer, which removes recurrence and convolution completely to rely purely on self-attention. The described design allows for paralleling and sharply reduces the number of training times compared to previous designs while improving the performance of the Translating Machine. The Transformer model involves both encoder and decoder, where encoder has 6 layers of multi-head mechanism as well as fully connected feed forward network and decoder also has similar elements. If scaled dot-product attention and multi-head attention will be successful, it means we are allowed to attend to different positions in the input sequence simultaneously and as such are good for parallelization and decreasing dependency path length. This is used where there is no recurrence and convolution to put order to sequences using positional encoding information. While translating from English to German in WMT 2014, Transformers translated words were ranked by means of the BLEU score of 28. For the Spanish to English task the NMT model obtains an ACL score of 4 whereas for the English to French task it has an ACL score of 41.

From these results, it can be concluded that the proposed model is better over conventional architectures in learning language representations and them interrelated. As with most neural networks, Transformer can be trained in parallel which again significantly reduces time required and resources needed. This efficiency is much more if the reached point is state – of – the – art level and the model becomes invaluable tool for the sequence transduction tasks. Preattention and Positional Encoding thus make the whole Transformer capable of processing assembled sequences of different lengths with very high accuracy and a very impressive processing speed.

In conclusion, the Transformers model elucidates that self-attention based approaches can be incorporated in lieu of RNNs and convolution where sequence transduction is required. There leads to evidence of the advantages in the training space and performance that supports this claim of this technique that transforms the Transformer into a remarkable machine translation and sequence modeling tool. The proposal of the Transformer represents an architectural innovation in constructing novel neural architectures; as a result, there exist broader potentials for expanding and deepening the domain of NLP. Such a transfer not only brings the potential for using the neural network for processing the sequential data into play but also opens a new direction to improving the training efficacy and accuracy in a number of tasks.