Generative AI

Name-Anmol Gupta
Semester-7th
Prn-21070521013

Attention all you need

In the paper “Attention Is All You Need,”, the authors propose a new elegant architecture of neural networks that can be very useful for certain tasks such as translation of text. Learning algorithms such as recurrent and convolutional neural networks served as the conventional approaches to the management of sequential data, albeit with some constraints. These older model fails to capture long range dependency and are slower because of their sequential nature. However, some modifications have been made to these models such as the addition of attention mechanisms it will still contain some of the problems mentioned above.

Methodology-
The paper introduces a new model of the NLP called the Transformer model, which liberated by employing only attention mechanisms. This architecture consists of an encoder and a decoder both of which are composed of elements of self-attention and elements of a point-wise fully connected networks. The third argument is the self-attention operation which gives the model a way of focusing in different positions in the input sequence, which is more efficient for creating long dependencies. A specific feature is called multi-head attention that provides the model with an opportunity to work with different positions and representational subspaces at the same time. Besides, this approach enhances the model’s ability in learning complicated sequences more efficiently and parallel computation.

Findings-
The Transformer model set new performance standards in language translation tasks. For instance, it achieved record-breaking BLEU scores of 28.4 for English-to-German and 41.8 for English-to-French translations on the WMT 2014 dataset. Moreover, it did this much faster than previous models, requiring only 3.5 days of training on eight GPUs for the English-to-French task. The model also showed strong generalization abilities, excelling in English constituency parsing with minimal adjustments.