**Attention Is All You Need**

***Authors* –** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser

***Summary:***

**Introduction**

* In the paper, the authors propose a new network called the Transformer, which is a network that was modelled using only attention mechanisms for sequence transduction problems.
* It synthesizes the original encoder-decoder architecture, where the recurrent or convolutional neural networks are used.

**1. Background**

* The paper gives an insight to the ‘Transformer’ which is a new network architecture, designed to handle sequence to sequence transduction tasks.
* It is based solely on attention mechanisms: Model Architecture has a unique focus on legal research, which this paper will attempt to demonstrate by analysing the work of researchers in this field.
* Elaborates on the architecture of the Transformer which has the encoder and decoder stacks, self-attention mechanism, and position-wise feed-forward layers and does not incorporate recurrent and convolutional archetypes.
* The Transformer returns more parallelizable, trains faster and results in better quality in-forward networks.

**2. Attention**

* Defines the Scaled Dot-Product Attention and Multi-Head Attention processes involved in the process of the Transformer in the machine translation task.
* RNNs and CNNs are used commonly.

**3. Training**

* Explains the training schedule of sequence modelling and transduction along with the data, batching, hardware, and optimizer.
* The attention mechanism has been incorporated with RNNs in all cases especially fade-in and fade-out cases.
* It also expounds the experiment results of machine translation, proving that the Transformer cannot merely outdo the previous models in translation but also establish itself as the first transduction model that depends solely on self-attention.

**4. Model Architecture**

* The Transformer has a quality of an encoder-decoder structure, yet, it has been found to undergo less training as per the best options observed in this study.

**5. Model Variations**

* Studies the modifications of the basic Transformer model and their changes with stacked self-attention and actual, point-wise fully connected ones.
* Both the encoder and decoder have many layers, including residual blocks and on performance.

**6. Conclusion**

* The paper describes the transformer, which is a new type of network architecture that is based on attention and surpasses recurrent or convolutional models in terms of quality in the translation of machine.
* The Transformer produces improved translation quality and is more efficient when it is closer to each position apart from the three mentioned networks.
* Positional encoding is applied to introduce additional information regarding the relative or absolute position of tokens which needs lower training time and can be parallelized.
* The paper also gives examples of applying the Transformer to other problems, for example, parse the English sentence. in the sequence.
* The Transformer does not have any challenge in achieving state of art accuracy in any of the machine translation tasks that have been performed on it, and this it does while reducing the training cost as compared to previous models.
* It generalizes well to other tasks such as English constituency parsing.
* The actual code used for training and their evaluation is provided for future work and modifications.

**By:**

**Name - Shreya Chaple**

**PRN - 21070521074**

**Section - B**