The paper namely “Attention Is All You Need” by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin present the Transformer, a new neural network architecture that eschews recurrence and convolutions in its entirety and exclusively employs self-attention. The described design enables more paralleling and reduces training time by several orders while providing better Translating Machine performance. The Transformer model contains an encoder and a decoder, each of which contains six layers of multi-head mechanism and fully connected feed forward network. Successful innovations like scaled dot-product attention and multi-head attention makes it possible to attend to different positions in the input sequence at the same time and thus are good for parallelization and decreasing dependency path length. This is applied in case of lack of recurrence and convolution to impose sequence order information known as positional encoding.

In the translation task of English to German in WMT 2014 the Transformer obtained the BLEU score of 28. On the Spanish to English task, the NMT model achieves an ACL score of 4, while on the English to French it performs with an ACL score of 41. 8, outperforming previous models. These results reveal that the proposed model is superior over traditional architectures in learning language representations and its dependencies. The Transformer can be trained in parallel which also greatly cuts down on time needed and computational power needed. This efficiency is even more impressive as the reached point is at the state- of – the- art level and the model proves to be a useful tool for the sequence transduction tasks. Preattention and Positional Encoding make the whole Transformer capable of dealing with assembled sequences at high accuracy and speed.

Summarising, in the Transformer model, it is proved that self-attention mechanisms can be used instead of classical RNNs and convolutions in sequence transduction. The benefits in the training space and performance support the claim of this technique, making the Transformer a breakthrough in the field of machine translation and sequence modeling. The introduction of the Transformer is an architectural advancement in the design of new neural networks, thus opening further possibilities of exploring and developing NLP in different domains. Such a transfer not only enriches the potential of using neural networks for processing the sequential data but also expands the perspectives of increasing the training effectiveness and accuracy in numerous tasks.