**SUMMARY-**

The “Attention Is All You Need” paper proposes an architecture known as the Transformer, which does not incorporate RNNs or CNNs into its structure but instead focuses on attention. The Transformer model is used for sequence transduction tasks such as machine translation and provides better training performance and accuracy than RNNs and CNNs. The major advancement of the Transformer model is the self-attention mechanism that permits parallelization and results in efficient training.

The Transformer model is composed of an encoder and a decoder which are made of multiple similar layers. The encoder transforms the input sequence into continuous representations that are then used by the decoder to produce the output sequence. Each layer in the encoder and decoder comprises two main components: A multi-head self-attention mechanism and a position-wise fully connected feed-forward network with residual connection and layer normalization. The self-attention mechanism allows the model to decide which parts of the input sequence should be used to compute each element of the output sequence, which makes it easier to capture long dependencies in the sequence compared to RNNs.

In experiments, the Transformer achieved results on the WMT 2014 English-to-German and English-to-French translation tasks, demonstrating its effectiveness and efficiency.