Attention mechanisms have been introduced, allowing models to focus on relevant parts of input sequences regardless of their distance but it is slow and memory-intensive for long sequences, so Transformers solve this problem with using self attention that allows for much greater parallelization, improving training speed and achieving state-of-the-art results in translation tasks. The encoder processes the input sequence using self-attention mechanisms that allow each word to attend to every other word in the sequence, capturing global context. Each encoder layer includes multi-head self-attention and a position-wise feed-forward neural network, with residual connections and layer normalization to stabilize training.

The decoder also consists of similar layers but includes an additional mechanism: masked self-attention to ensure that predictions are made one token at a time without looking ahead, preserving the sequence generation nature of the task. It also incorporates encoder-decoder attention, which enables the decoder to focus on relevant parts of the input sequence when generating output.

Both input and output tokens are first passed through learned embeddings and positional encodings to retain information about word order, which the model otherwise lacks due to its non-recurrent design. The final output of the decoder is passed through a linear layer followed by a softmax function to produce the predicted tokens. This design not only allows for highly parallel computation but also improves performance on long sequences by modeling global dependencies efficiently.