**A summary on Attention is all we need**

In this paper, the Transformer model has been presented by the research team in charge with the work “Attention Is All You Need” as a revolutionary architecture that is used to complete sequence transduction tasks, and is especially suitable for language translation. It is on attention mechanisms that Transformer relies on offering an architecture that does not require recurrence or convolutional layers, which greatly facilitates training parallelization and reduces sequence processing time. The overall framework of the Transformer model is divided into two key segments: the encoder and the decoder. While the encoder transforms the input sequence into continuous representations, the decoder transforms those representations into the output sequence. The two, encoder, as well as decoder, take advantage of self-attention that makes it easier for the model to see all parts of an input at any given time during prediction. This design helps the Transformer understand complex relationships in the data, making it very effective for tasks like translation. A key feature is its self-attention mechanism, enabling the model to assess the significance of different words in a sentence during encoding, effectively capturing long-range dependencies. The use of multi-head attention allows the model to concurrently focus on information from various representation subspaces at different positions, enhancing its ability to understand context and relationships within the data. Since the Transformer lacks an inherent sense of sequence order, it uses positional encodings to provide information about the position of words in a sequence. The model achieves state-of-the-art results on several benchmark translation tasks, including the WMT 2014 English-to-German and English-to-French translation tasks, surpassing previous models and being more efficient in training time. Its highly scalable architecture allows for training very large models that can handle extensive datasets effectively. Although the paper primarily focuses on translation tasks, the Transformer architecture has since been applied to various other NLP tasks, such as text summarization and question answering. Overall, the paper demonstrates that attention mechanisms alone can achieve superior performance in sequence transduction tasks, paving the way for future research and applications in deep learning and natural language processing.