**Name: Parth Sunil Satokar**

**Sec: B**

**PRN: 21070521113**

**CA1 Generative AI**

**Summary**

The paper introduces the Transformer, a novel architecture for sequence transduction tasks, which relies entirely on self-attention mechanisms, eliminating the need for recurrence and convolutions. Traditional models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs) have been the state-of-the-art for these tasks, but they face challenges with parallelization due to their sequential nature. The Transformer addresses these issues by allowing more parallelization and reducing training time.

The Transformer model consists of an encoder and a decoder, both composed of stacked layers. Each encoder layer has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The decoder layer adds a third sub-layer, performing multi-head attention over the encoder’s output. Residual connections and layer normalization are applied around each sub-layer.

The core component of the Transformer is the attention mechanism, particularly the scaled dot-product attention. This mechanism computes attention scores by taking the dot product of query and key vectors, scaling them, and applying a SoftMax function to obtain the weights. The Transformer employs multi-head attention, allowing the model to jointly attend to information from different representation subspaces.

The Transformer demonstrates superior performance in machine translation tasks. On the WMT 2014 English-to-German translation task, it achieves a BLEU score of 28.4, surpassing previous models by over 2 BLEU points. For the WMT 2014 English-to-French task, it achieves a state-of-the-art BLEU score of 41.8. The Transformer’s architecture allows it to train significantly faster, completing in 3.5 days on eight GPUs compared to much longer times for traditional models.

The Transformer model represents a significant advancement in sequence transduction, offering better performance and efficiency. By leveraging self-attention and eliminating sequential dependencies, it opens up new possibilities for parallelization and faster training times, making it a valuable model for various natural language processing tasks. The paper's results highlight the potential of self-attention mechanisms to replace more complex recurrent and convolutional structures in neural network models.