Attention Is All You Need

This paper introduces the **Transformer** model – a novel neural network architecture for sequence transduction tasks like machine translation. Unlike traditional models that depend on **recurrent (RNN) or convolutional (CNN)** layers, the Transformer relies **entirely on attention mechanisms**, specifically **multi-head self-attention**.

The Transformer enables **parallel processing**, which dramatically speeds up training. It achieved **state-of-the-art BLEU scores** on:

- **English→German** (WMT 2014): 28.4 BLEU (2 points better than best existing models, including ensembles).
- **English→French** (WMT 2014): 41.8 BLEU in just 3.5 days using 8 GPUs.

The architecture handles **long-range dependencies** better and generalizes well to other tasks like **English constituency parsing**. This work laid the foundation for later models such as **BERT** and **GPT**

Summary:

- **RNNs** struggle with long sequences and are slow to train.
- **CNNs** improve speed and parallelism but still lack full context awareness.
- **Transformers** revolutionized NLP by removing recurrence, enabling full parallelism, and outperforming both in quality and efficiency.