Summary of the Paper

"Attention Is All You Need"

Link:- https://arxiv.org/pdf/1706.03762

Problem:-

Traditional sequence-to-sequence models (like RNNs, LSTMs, GRUs) are effective for tasks such as machine translation, but they:

- Process sequences sequentially, step by step.
- Cannot be parallelized easily, making them slow to train and limiting scalability.
- Struggle with long-range dependencies.

Solution:-

- Removes recurrence completely.
- Uses self-attention mechanisms instead of RNNs or CNNs.
- Can process entire sequences in parallel, making training faster and more efficient.
- Achieves better translation quality than existing methods.

Results:-

- The Transformer outperforms previous state-of-the-art models in English-to-German and English-to-French translation tasks.
- Training takes less time (e.g., 12 hours on 8 GPUs).
- It sets a new benchmark for sequence transduction tasks.

Encoder:-

The encoder maps an input sequence of symbol representations (x1, ..., xn) to a sequence of continuous representations z = (z1, ..., zn). Given z.

Decoder:-

The decoder then generates an output sequence (y1, ..., ym) of symbols one element at a time. At each step the model is auto regressive consuming the previously generated symbols as additional input when generating the next.

Self-attention:-

Self-attention is a mechanism that allows a model to weigh the importance of different words in a sequence when processing each word. For every word in the input, self-attention calculates how much attention should be paid to other words in the same sequence.

Positional encoding:-

Since the Transformer has no recurrence or convolution, it has no natural way to understand the order of words in a sequence. Positional encoding solves this by adding special vectors to the input embeddings that contain information about the position of each word. These encodings are generated using sinusoidal functions, allowing the model to infer both absolute and relative positions.