"Attention Is All You Need" (Transformer Architecture)

Attention mechanisms enable models to dynamically focus on relevant parts of input sequences, regardless of distance, but traditional approaches (e.g., RNNs with attention) are slow and memory-intensive for long sequences. The **Transformer** solves these limitations by introducing **self-attention**, which processes all positions in parallel, drastically improving training speed while achieving state-of-the-art results in translation tasks.

Encoder: Capturing Global Context

The encoder processes the input sequence using **self-attention mechanisms**, where each word attends to every other word in the sequence, capturing full **global dependencies**. Each encoder layer consists of:

1. **Multi-head self-attention** – Computes attention across multiple representation subspaces, allowing the model to jointly focus on different positional and semantic relationships.
2. **Position-wise feed-forward network (FFN)** – Applies the same fully connected network to each token independently.
3. **Residual connections + layer normalization** – Stabilizes training by mitigating gradient issues in deep networks.

Decoder: Autoregressive Sequence Generation

The decoder shares a similar structure but includes two critical modifications for autoregressive prediction:

1. **Masked self-attention** – Prevents the model from "cheating" by attending to future tokens, ensuring predictions depend only on previously generated outputs.
2. **Encoder-decoder attention** – Lets the decoder focus on relevant parts of the input sequence (from the encoder’s output) while generating each token.

Token Representation & Positional Encoding

Since the Transformer lacks recurrence, it must explicitly encode sequential information:

- **Learned embeddings** convert input/output tokens to vectors.
- **Sinusoidal positional encodings** are added to embeddings to preserve word order, using fixed geometric patterns (sine/cosine functions) that generalize to unseen sequence lengths.

Output Generation

The decoder's final output passes through:

1. A **linear projection layer** to map embeddings to vocabulary space.
2. A **softmax activation** to produce token probabilities.

Key Advantages

- **Parallelization**: Self-attention eliminates sequential dependencies, enabling faster training than RNNs.
- **Long-range dependency modeling**: Global attention captures relationships between distant tokens more effectively than recurrent or convolutional methods.
- **Scalability**: The architecture's efficiency allows scaling to deeper/larger models (later exploited by BERT, GPT, etc.).

The Transformer became the foundation for modern NLP, outperforming RNN/CNN-based models in translation (e.g., +2 BLEU on WMT 2014 English-German) while reducing training time. Its design principles now underpin nearly all state-of-the-art language models.