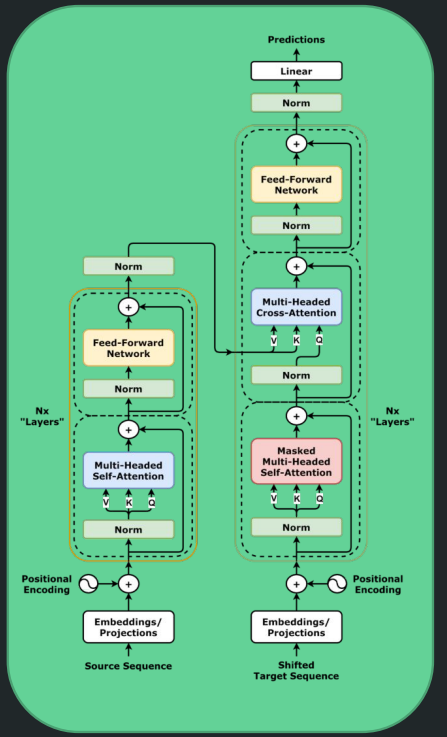
2.3. Transformer Architecture

Transformers

Revolutionizing Natural Language Processing

- Process entire sentences at once (not word-by-word).
- Powered by self-attention → captures word importance & context.
- Origin: “Attention Is All You Need” (2017).
- Excels at translation, text generation, and more.



Core Architecture

Encoder

The encoder converts input text into rich contextual embeddings through multiple stacked layers. Each layer combines self-attention and feed-forward networks, enabling bidirectional processing for complete context understanding.

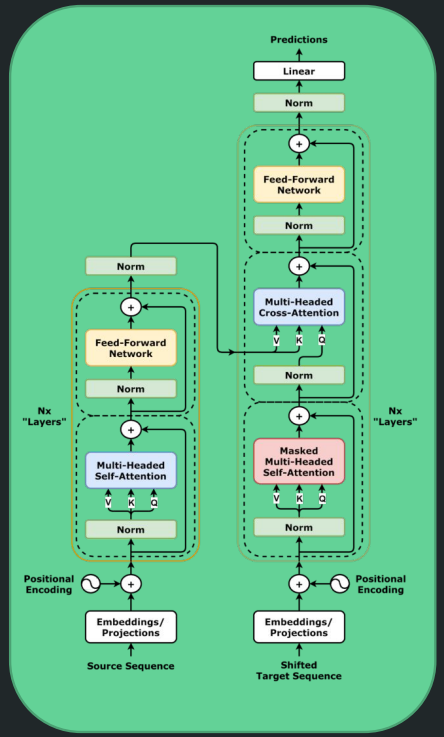
- Transforms input into contextual embeddings
- Stacked layers with self-attention + feed-forward
- Bidirectional for full context capture

Decoder

The decoder generates output sequences step by step, using masked self-attention to preserve causality and cross-attention to integrate encoder information, ensuring coherent and contextually accurate outputs.

- Produces output auto-regressively (one token at a time)
- Masked self-attention prevents future-token leakage
- Cross-attention leverages encoder representations

Self-Attention Mechanism



Purpose

Allows the model to weigh the importance of each word in a sequence relative to all others, enabling rich context understanding.

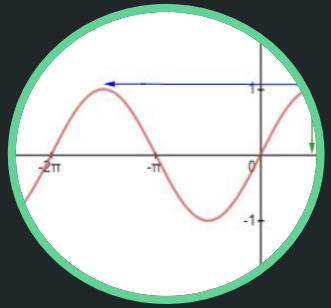
Query–Key–Value

Self-attention uses Query, Key, and Value vectors (from input embeddings via linear projections) to model relationships between tokens.

Attention Computation

Attention scores = similarity between Query and Key, normalized with softmax, then applied to Values ⟶ produces context-aware representations for each token.

Self-Attention Mechanism



Positional Encoding

Transformers look at all words at once, so they need a way to know the order of words in a sentence. Positional encoding adds this order information. Using patterns like sine and cosine waves, each word gets a unique “position signal.” When added to word embeddings, the model can understand both the meaning of words and their order.

Training Process

Transformers are first trained on huge amounts of text (pre-training) and then adjusted for specific tasks (fine-tuning). They learn in a self-supervised way, using methods like filling in missing words (masked language modeling) or predicting the next sentence. By seeing billions of words, they pick up grammar, meaning, and patterns in language, which makes them useful for many applications.