Transformers and LLMs

Suryansh Sanatan

December 2025

1 What is a transformer

Transformer is a very useful type of neural network used in the modern day LLMs with major use cases involving predicting next word and translation.

1.1 Encoder

first part of a transformer is encoder.it basically generates embedding of each word in the sentence keeping the context in mind also there are beginning of sentence **BOS** and end of sentence **EOS** tokens.one common model used based on an encoder is BERT(developed by google)

first it tokenizes the sentence then it adds a positional embeddings(PE)token of position to each of the tokens as this processes the whole sentence at a time.

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right)$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right)$$

This formula states the positional embeddings for odd(2i+1) and even(2i) dimension indexes(i) for word at the *pos index*.

1.1.1 Attention

now the new embeddings are passed through a neural network to generate **Query Matrix** (which is basically like asking a question) which is compared with other words **Key Matrices**(like answer to the question which is also generated by a different neural network) by dot product **Similarity score** is generated with respect to all other words which by softmax function is converted to numbers between 0-1

now using a different neural network we create **Value Matrix** after this we multiply this with similarity scores(after softmax) and *voila !* we have generated the contextual embeddings by adding self attention values to initial embeddings

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\intercal}}{\sqrt{d_k}}\right)V$$

Here, d_k is the k^{th} dimension of the keys.

1.2 Decoder

decoder basically takes the input from encoder and predicts the next word or translation of the sentence.one model based on decoding is GPT.it follows an architecture similar to the encoder just there are masked cross attention heads in this that help to encode the context by computing dot of query and key generated by encoder with value of the right shifted outputs to to encode context of input to the output

1.3 Multilayer Perceptron

This helps to encode facts into the LLM like if we input michael jordan it should output basketball.

2 Large Language models(LLMs)

so a large language model has multiple sandwitched attention blocks and multilayer perceptron that train at a very high speed and are highly effective due to the highly parallisable using GPUs that has really helped in the developmet of tools like Gemini Chatgpt