TRANSFORMER MODEL (USED IN CHAT GPT)

GPT stands for “Generative Pre-trained Transformer” which is a type of neural network architecture specially designed for natural language processing. It is capable of generating new content and the model undergoes a pre training period where it’s exposed to vast amount of data allowing it replicate and generate data.

The Transformer model is the underlying architecture model of GPT. Generative pretrained transformers are a family of Transformer models trained by OpenAI for Language Modelling tasks.

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. It is one of the latest forms of deep learning, they apply an evolving set of mathematical techniques, called self-attention where the model looks at an input sequence and decides which parts of the sequence are important and adds weight to different words or elements of that particular sequence. Transformers use multiple sets of self-attention mechanisms, known as attention heads. This allows the model to learn different aspects of the relationships between words in parallel. The math that transformers use lends itself to parallel processing, so these models can run fast, because of this the transformers don’t really understand the order of a sequence (e.g., words in a sentence) since they process all positions in parallel. To give the model information about the sequence's order, positional encoding is added to the input embeddings. After attention mechanisms, the transformer model typically includes feedforward neural networks to process and transform the information. Feed forward neural networks are artificial neural networks where the nodes do not form loops, they are also known as multi-layer neural layer network as all the information is passed forward. Finally, to speed up and stabilize the training process Layer Normalization and Residual Connections are used.

Transformer is an architecture for transforming one sequence into another one with the help of two parts (Encoder and Decoder), but it differs from the previously existing sequence-to-sequence models because it does not use any Recurrent Networks. Sequence-to-Sequence (or Seq2Seq) is a neural net that transforms a given sequence of elements.

Before transformers arrived, users had to train neural networks with large, labelled datasets that were costly and time-consuming to produce. By finding patterns between elements mathematically, transformers eliminate that need, making available the trillions of images and petabytes of text data on the web and in corporate database. They translate text and speech in near real-time, opening meetings and classrooms to diverse and hearing-impaired attendees. Like most neural networks, transformer models are basically large encoder/decoder blocks that process data. Transformers use positional encoders to tag data elements coming in and out of the network. Attention units follow these tags, calculating a kind of algebraic map of how each element relates to the others. Attention queries are typically executed in parallel by calculating a matrix of equations in what’s called multi-headed attention. This allows the transformers to find patters similar to a human and generate the desired outputs. These models can generate relevant and desired outputs but they lack consciousness or understanding. They operate purely on patterns and sequences learned during the training era.