Explained | What is a transformer, the ML model that powers ChatGPT?

Makarand Tapaswi

Machine learning (ML), a subfield of artificial intelligence, teaches computers to solve tasks based on structured data, language, audio, or images, by providing examples of inputs and the desired outputs. This is different from traditional computer programming, where programmers write a sequence of specific instructions. Here, the ML model *learns* to generate desirable outputs by adjusting its many knobs – often in the millions.

ML has a history of developing methods with hand-crafted features that may work only for specific, narrow problems. There are several such examples. In text, classifying a document as scientific or literary may be solved by counting the number of times certain words appear. In audio, spoken text is recognised by converting the audio into a time-frequency representation. In images, a car may be found by checking for the existence of specific car-like edge-shaped patterns.

Such hand-crafted features are combined with simple, or shallow, learning classifiers that typically have up to tens of thousands of knobs. In technical parlance, these knobs are called parameters.

Deep neural networks

In the first part of the 2010s, deep neural networks (DNNs) took over ML by storm, replacing the classic pipeline of hand-crafted features and simple classifiers. DNNs ingest a complete document or image and generate a final output, without the need to specify a particular way of extracting features.

While these deep and large models have existed in the past, their large size – millions of parameters – hindered their use. The resurgence of DNNs in the 2010s is attributed to the availability of large-scale data and fast parallel computing chips called graphics processing units.

Further, the models used for text or images were still different: recurrent neural networks were popular in language-understanding while convolutional neural networks (CNNs) were popular in computer vision, i.e. machine understanding of the visual world.

‘ Attention Is All You Need’

In a pioneering paper entitled ‘Attention Is All You Need’ that appeared in 2017, a team at Google proposed transformers – a DNN architecture that has today gained popularity across all modalities: image, audio, and language. The original paper proposed transformers for the task of translating a sentence from one language to another, similar to what Google Translate does when converting from, say, English to Hindi.

A transformer is a two-part neural network. The first part is an ‘encoder’ that ingests the input sentence in the source language (e.g. English); the second is a ‘decoder’ that generates the translated sentence in the target language (Hindi).

The encoder converts each word in the source sentence to an abstract numerical form that captures the meaning of the word within the context of the sentence, and stores it in a memory bank. Just like a person would write or speak, the decoder generates one word at a time referring to what has been generated so far and by looking back

at the memory bank to find the appropriate word. Both these processes use a mechanism called ‘attention’, hence the name of the paper.

A key improvement over previous methods is the ability of a transformer to translate long sentences or paragraphs correctly.

The adoption of transformers subsequently exploded. The capital ‘T’ in ChatGPT, for example, stands for ‘transformer’.

Transformers have also become popular in computer vision: they simply cut an image into small square patches and line them up, just like words in a sentence. By doing so, and after training on large amounts of data, a transformer can provide better performance than CNNs. Today, transformer models constitute the best approach for image classification, object detection and segmentation, action recognition, and a host of other tasks.

Transformers’ ability to ingest anything has been exploited to create joint vision-and-language models that allow users to search for an image (e.g. Google Image Search), describe one, and even answer questions regarding the image.

What is ‘attention’?

Attention in ML allows a model to learn how much importance should be given to different inputs. In the translation example, attention allows the model to select or weigh words from the memory bank when deciding which word to generate next. While describing an image, attention allows models to look at the relevant parts of the image when generating the next word.

A fascinating aspect of attention-based models is their ability for self-discovery, by parsing a lot of data. In the translation case, the model is never told that the word “dog” in English means “कुत्ता” in Hindi. Instead, it finds these associations by seeing several training sentence pairs where “dog” and “कुत्ता” appear together.

A similar observation applies to image captioning. For an image of a “bird flying above water”, the model is never told which region of the image corresponds to “bird” and which “water”. Instead, by training on several image-caption pairs with the word “bird”, it discovers common patterns in the image to associate the flying thing with “bird”.

Transformers are attention models on steroids. They feature several attention layers both within the encoder, to provide meaningful context across the input sentence or image, and from the decoder to the encoder when generating a translated sentence or describing an image.

The billion and trillion scale

In the last year, transformer models have become larger and train on more data than before. When these colossuses train on written text, they are called large language models (LLMs). ChatGPT uses hundreds of billions of parameters whereas GPT-4 uses hundreds of trillions.

While these models are trained on simple tasks, such as filling in the blanks or predicting the next word, they are very good at answering questions, creating stories, summarising documents, writing code, and even solving mathematical word problems in steps. Transformers are also the bedrock of *generative* models that create realistic images and audio. Their utility in diverse domains makes transformers a very powerful and universal model.

However, there are some concerns. The scientific community is yet to figure out how to evaluate these models rigorously. There are also instances of “hallucination”, whereby models make confident but wrong claims. We must urgently address societal concerns, such as data privacy and attribution to creative work, that arise as a result of their use.

At the same time, given the tremendous progress, ongoing efforts to create guardrails guiding their use, and work on leveraging these models for positive outcomes (e.g. in healthcare, education, and agriculture), optimism wouldn’t be misplaced.

Dr. Makarand Tapaswi is a senior machine learning scientist at Wadhwani AI, a non-profit on using AI for social good, and an assistant professor at the computer vision group at IIIT Hyderabad, India.