Google's Gemini: Is the new AI model really better than ChatGPT?

December 15 2023, by Michael G. Madden

Credit: Pixabay/CC0 Public Domain

Google Deepmind has recently announced Gemini, its new AI model to compete with OpenAI's ChatGPT. While both models are examples of "generative AI," which learn to find patterns of input training information to generate new data (pictures, words or other media), ChatGPT is a large language model (LLM) which focuses on producing text.

In the same way that ChatGPT is a web app for conversations that is based on the neural network know as GPT (trained on huge amounts of text), Google has a conversational web app called Bard which was based on a model called LaMDA (trained on dialogue). But Google is now upgrading that based on Gemini.

What distinguishes Gemini from earlier generative AI models such as LaMDA is that it's a "multi-modal model." This means that it works directly with multiple modes of input and output: as well as supporting text input and output, it supports images, audio and video. Accordingly, a new acronym is emerging: LMM (large multimodal model), not to be confused with LLM.

In September, OpenAI announced a model called GPT-4Vision that can work with images, audio and text as well. However, it is not a fully multimodal model in the way that Gemini promises to be.

For example, while ChatGPT-4, which is powered by GPT-4V, can work with audio inputs and generate speech outputs, OpenAI has confirmed that this is done by converting speech to text on input using another deep learning model called Whisper. ChatGPT-4 also converts text to speech on output using a different model, meaning that GPT-4V itself is working purely with text.

Likewise, ChatGPT-4 can produce images, but it does so by generating text prompts that are passed to a separate deep learning model called Dall-E 2, which converts text descriptions into images.

In contrast, Google designed Gemini to be "natively multimodal." This means that the core model directly handles a range of input types (audio, images, video and text) and can directly output them too.

The verdict

The distinction between these two approaches might seem academic, but it's important. The general conclusion from Google's technical report and other qualitative tests to date is that the current publicly available version of Gemini, called Gemini 1.0 Pro, is not generally as good as GPT-4, and is more similar in its capabilities to GPT 3.5.

Google also announced a more powerful version of Gemini, called Gemini 1.0 Ultra, and presented some results showing that it is more powerful than GPT-4. However, it is difficult to assess this, for two reasons. The first reason is that Google has not released Ultra yet, so results cannot be independently validated at present.

The second reason why it's hard to assess Google's claims is that it chose to release a somewhat deceptive demonstration video, see below. The video shows the Gemini model commenting interactively and fluidly on a live video stream.

However, as initially reported by Bloomberg, the demonstration in the video was not carried out in real time. For example, the model had learned some specific tasks beforehand, such the three cup and ball trick, where Gemini tracks which cup the ball is under. To do this, it had been provided with a sequence of still images in which the presenter's hands are on the cups being swapped.

Promising future

Despite these issues, I believe that Gemini and large multimodal models are an extremely exciting step forward for generative AI. That's both because of their future capabilities, and for the competitive landscape of AI tools. As I noted in a previous article, GPT-4 was trained on about 500 billion words—essentially all good-quality, publicly available text.

The performance of deep learning models is generally driven by

increasing model complexity and amount of training data. This has led to the question of how further improvements could be achieved, since we have almost run out of new training data for language models. However, multimodal models open up enormous new reserves of training data—in the form of images, audio and videos.

AIs such as Gemini, which can be directly trained on all of this data, are likely to have much greater capabilities going forward. For example, I would expect that models trained on video will develop sophisticated internal representations of what is called "naïve physics." This is the basic understanding humans and animals have about causality, movement, gravity and other physical phenomena.

I am also excited about what this means for the competitive landscape of AI. For the past year, despite the emergence of many generative AI models, OpenAI's GPT models have been dominant, demonstrating a level of performance that other models have not been able to approach.

Google's Gemini signals the emergence of a major competitor that will help to drive the field forward. Of course, OpenAI is almost certainly working on GPT-5, and we can expect that it will also be multimodal and will demonstrate remarkable new capabilities.

All that being said, I am keen the see the emergence of very large multimodal models that are open-source and non-commercial, which I hope are on the way in the coming years.

I also like some features of Gemini's implementation. For example, Google has announced a version called Gemini Nano, that is much more lightweight and capable of running directly on mobile phones.

Lightweight models like this reduce the environmental impact of AI computing and have many benefits from a privacy perspective, and I am

sure that this development will lead to competitors following suit.

Provided by The Conversation