Meta unveils a new large language model that can run on a single GPU

Benj Edwards / Ars Technica

reader comments

44 with

Share this story

On Friday, Meta announced a new AI-powered large language model (LLM) called LLaMA-13B that it claims can outperform OpenAI's GPT-3 model despite being "10x smaller." Smaller-sized AI models could lead to running ChatGPT-style language assistants locally on devices such as PCs and smartphones. It's part of a new family of language models called "Large Language Model Meta AI," or LLAMA for short.

The LLaMA collection of language models range from 7 billion to 65 billion parameters in size. By comparison, OpenAI's GPT-3 model—the foundational model behind ChatGPT—has 175 billion parameters.

Meta trained its LLaMA models using publicly available datasets, such as Common Crawl, Wikipedia, and C4, which means the firm can potentially release the model and the weights open source. That's a dramatic new development in an industry where, up until now, the Big Tech players in the AI race have kept their most powerful AI technology to themselves.

"Unlike Chinchilla, PaLM, or GPT-3, we only use datasets

publicly available, making our work compatible with open-sourcing and reproducible, while most existing models rely on data which is either not publicly available or undocumented," tweeted project member Guillaume Lample.

> *Today we release LLaMA, 4 foundation models ranging from 7B to 65B parameters.*
> *LLaMA-13B outperforms OPT and GPT-3 175B on most benchmarks.*
> *LLaMA-65B is competitive with Chinchilla 70B and PaLM 540B.*
> *The weights for all models are open and available at https://t.co/q51f2oPZlE*
> *1/n pic.twitter.com/DPyJFBfWEq*
>
> *— Guillaume Lample (@GuillaumeLample) February 24, 2023*

Meta calls its LLaMA models "foundational models," which means the firm intends the models to form the basis of future, more-refined AI models built off the technology, similar to how OpenAI built ChatGPT from a foundation of GPT-3. The company hopes that LLaMA will be useful in natural language research and potentially power applications such as "question answering, natural language understanding or reading comprehension, understanding capabilities and limitations of current language models."

Advertisement

While the top-of-the-line LLaMA model (LLaMA-65B, with 65 billion parameters) goes toe-to-toe with similar offerings from competing AI labs DeepMind, Google, and OpenAI, arguably the most interesting development comes from the LLaMA-13B model, which, as previously mentioned, can reportedly outperform GPT-3 while running on a single GPU. Unlike the data center requirements for GPT-3 derivatives, LLaMA-13B opens the door for ChatGPT-like performance on consumer-level hardware in the near future.

Parameter size is a big deal in AI. A parameter is a variable that a machine-learning model uses to make predictions or

classifications based on input data. The number of parameters in a language model is a key factor in its performance, with larger models generally capable of handling more complex tasks and producing more coherent output. More parameters take up more space, however, and require more computing resources to run. So if a model can achieve the same results as another model with fewer parameters, it represents a significant gain in efficiency.

“I’m now thinking that we will be running language models with a sizable portion of the capabilities of ChatGPT on our own (top of the range) mobile phones and laptops within a year or two,” wrote independent AI researcher Simon Willison in a Mastodon thread analyzing the impact of Meta’s new AI models.

Currently, a stripped-down version of LLaMA is available on GitHub. To receive the full code and weights (the “learned” training data in a neural network), Meta provides a form where interested researchers can request access. Meta has not announced plans for a wider release of the model and weights at this time.

https://arstechnica.com/?p=1919928