LLaMA: Open and Efficient Foundation Language Models

⚠ This section is under heavy development.

What's new?

This paper introduces a collection of foundation language models ranging from 7B to 65B parameters.

The models are trained on trillion of tokens with publicly available datasets.

The work by (Hoffman et al. 2022) shows that given a compute budget smaller models trained on a lot more data can achieve better performance than the larger counterparts. This work recommends training 10B models on 200B tokens. However, the LLaMA paper finds that the performance of a 7B model continues to improve even after 1T tokens.

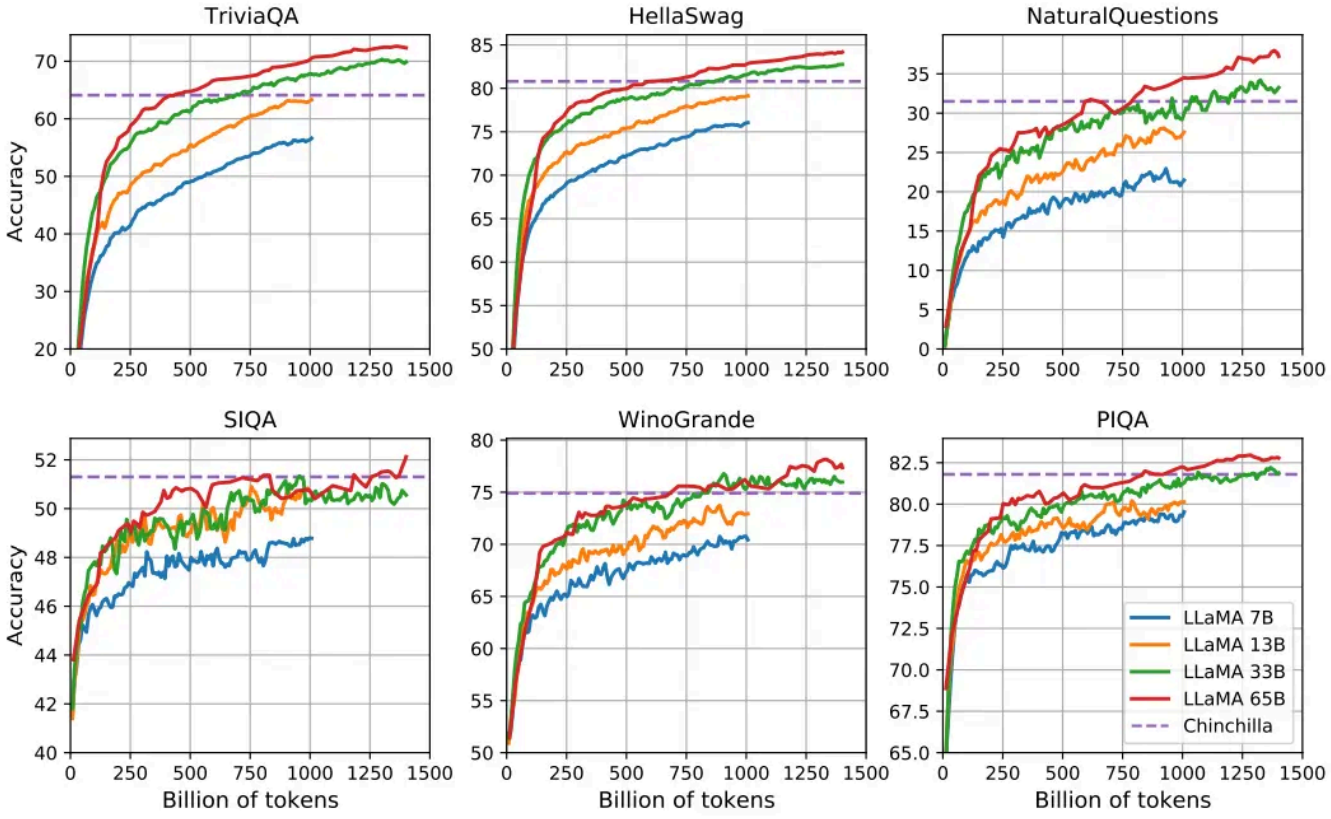

Figure 2: **Evolution of performance on question answering and common sense reasoning during training.**

This work focuses on training models (LLaMA) that achieve the best possible performance at various inference budgets, by training on more tokens.

Capabilities & Key Results

Overall, LLaMA-13B outperform GPT-3(175B) on many benchmarks despite being 10x smaller and possible to run a single GPU. LLaMA 65B is competitive with models like Chinchilla-70B and PaLM-540B.

Paper: LLaMA: Open and Efficient Foundation Language Models

Code: https://github.com/facebookresearch/llama

References

- Koala: A Dialogue Model for Academic Research (April 2023)
- Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data (April 2023)
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (March 2023)
- LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention (March 2023)
- GPT4All (March 2023)
- ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge (March 2023)
- Stanford Alpaca (March 2023)

Last updated on September 19, 2024