A short summary of open source LLMs

Summary:

During the last few months, several open source LLMs have been released [1]. In this note, we define "open source LLM" as a model that shares the model architecture and the dataset used publicly, but this does not mean that pretrained or fine-tuned models are also released.

With these open source models and datasets, many people can fine-tune manageable size LLMs (~10B) with their own or open source datasets. The models with this size range can run from a laptop or single server machine by using quantization or LoRA[7] to reduce the model size and run time overhead.

In addition to fine-tuned models, there are also models that are pretrained from scratch, and can be used commercially since pretrained llama [2] is not allowed for commercial use.

These advancements from open source are expected to continue and accelerate as more people will be participating.

People now know how 'a large hammer (100B or larger LLM)' can be useful, and want to use more of it. However, people also realized and try to use 'a small hammer' instead for different tasks to reduce cost.

Pretrained model, meta's llama

Meta has released Llama[2], pretrained models (7, 13, 33, 16B parameters), which are claimed to be similar to GPT-3. What's important with Llama is that the models with smaller sizes can achieve high performance as much as larger models (GPT-3, 175B) by training with large datasets. In terms of model architecture is very similar to GPT-2[11], decoder model based on Transformer.

Fine-tuning llama

Since the llama's release, several developments have occurred in the following ways:

1. fine-tuned llama model with instruction-following fine-tuning to be capable of chat assistant
2. created a dataset that allows the instruction-following fine-tuning by using existing LLM (ChatGPT) directly or indirectly (ShareGPT [5])

We found these are important for the following reasons:

- more people are now able to fine-tune manageable size (~10B or less) pretrained LLM to fine-tune the model to their own need. Fine-tuning cost from the recent models are around $1,000 or less range.
- more people are able to generate better dataset (instruction-following) which is one of the key ingredients to create chatGPT style LLM

Stanford's Alpaca[3]:

- fine-tuned llama 7B model using instruction-following method, cost $50
- generated dataset using GPT-3.5 (text-davinci-003), 52K, cost $500

Vicuna 13B [4]:

- fine-tuned llama 13B model, cost $300
- generated dataset using ShareGPT (site that shares chatgpt conversation) 70K dataset

GPT4ALL [6]:

- fine-tuned llama 7B, cost including all experiments ~ $900
- generated dataset using GPT-3.5 (text-davinci-003), 800K, cost $500
- used 4bit quantization
- LoRA(low-rank adaptation) and CPP

Pretrained model

Since llama license does not allow commercial usage, there are models that are pretrained from scratch so that they can be used commercially. Note that a model architecture can still be based on llama, but a model in this category pretrains from scratch with their own dataset.

MPT (mosaic pretrained transformer)[8]:

- model based on llama-7B, pretrain cost ~$200K
- can be used commercially
- context size is 65K, note that GPT-4's context size is ~40K

Openllama (open reproduction of llama)[9]:

- llama 7B model
- RedPajama dataset [10] is used

References:

1. Leaked google researcher's post,Google *We Have No Moat, And Neither Does OpenAI*
2. meta's llama
3. Stanford's alpaca
4. Vicuna 13B
5. ShareGPT
6. GPT4all
7. LoRa, low-rank adaptation
8. MPT

9. Openllama
10. RedPajama dataset
11. GPT-2 code