

LLM Fine-tuning

Nasy

May 19, 2023

Outline




- ① Introduction
- ② How?
- ③ Conclusion
- ④ Reference
- ⑤ Examples

Introduction

- Which model to choice?
- How to fine-tune?
- Examples

Which

Table 1. Elo ratings of LLMs (Timeframe: April 24 - May 8, 2023)

Rank	Model	Elo Rating	Description	License
1	 GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	 Claude-v1	1224	Claude by Anthropic	Proprietary
3	 GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Weights available; Non-commercial
5	Koala-13B	1022	a dialogue model for academic research by BAIR	Weights available; Non-commercial
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance	Apache 2.0
7	Oasst-Pythia-12B	928	an Open Assistant for everyone by LAION	Apache 2.0
8	ChatGLM-6B	918	an open bilingual dialogue language model by Tsinghua University	Weights available; Non-commercial
9	StableLM-Tuned-Alpha-7B	906	Stability AI language models	CC-BY-NC-SA-4.0
10	Alpaca-13B	904	a model fine-tuned from LLaMA on instruction-following conversations by OpenAI	Weights available; Non-commercial

LLaMA: Open and Efficient Foundation Language Models³

- The same model and architecture as GPT-2
 - Replace ReLU with SwiGLU [PaLM]¹
 - Rotary Embeddings [GPTNeo]²
- Publicly available datasets
- 7B, 13B, 33B(30B?), 65B

¹Noam Shazeer. *GLU Variants Improve Transformer*. Feb. 12, 2020. DOI: [10.48550/arXiv.2002.05202](https://doi.org/10.48550/arXiv.2002.05202). arXiv: 2002.05202 [cs, stat]. URL: <http://arxiv.org/abs/2002.05202> (visited on 05/19/2023). preprint.

²Jianlin Su et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. Aug. 8, 2022. DOI: [10.48550/arXiv.2104.09864](https://doi.org/10.48550/arXiv.2104.09864). arXiv: 2104.09864 [cs]. URL: <http://arxiv.org/abs/2104.09864> (visited on 05/19/2023). preprint.

³Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. Feb. 27, 2023. arXiv: 2302.13971 [cs]. URL: <http://arxiv.org/abs/2302.13971> (visited on 05/19/2023). preprint.

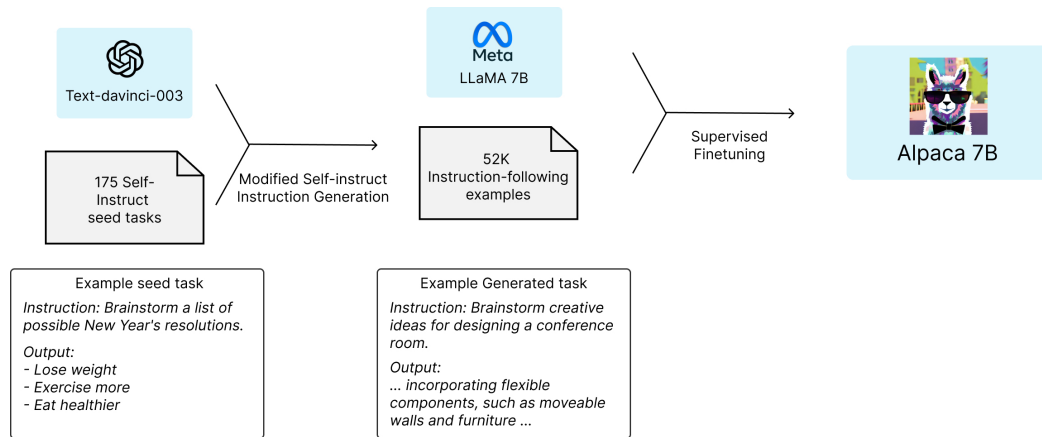
LLaMA

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Datasets

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

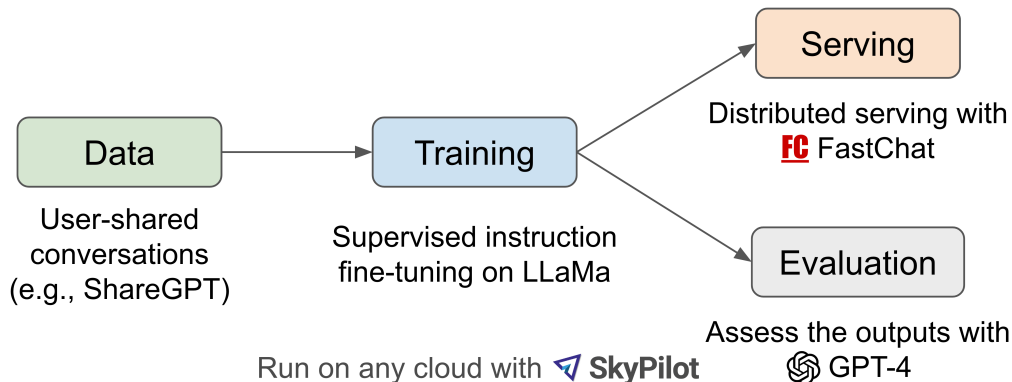
Alpaca



Alpaca

```
[{  
  "instruction": "Rewrite the following sentence in the third person",  
  "input": "I am anxious",  
  "output": "She is anxious."  
},  
{  
  "instruction": "What are the three primary colors?",  
  "input": "",  
  "output": "The three primary colors are red, blue, and yellow."  
}]
```

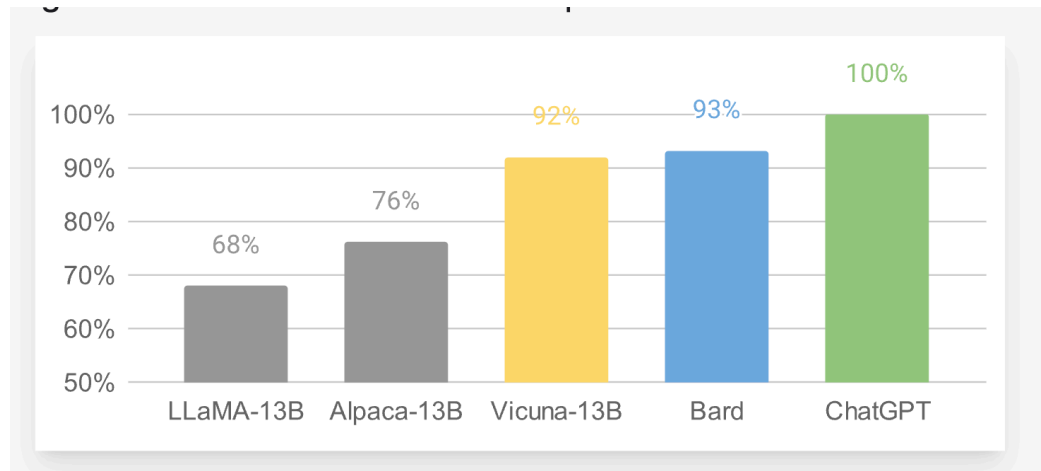
Vicuna



Vicuna

```
[
  {
    "from": "human",
    "value": "Who are you?"
  },
  {
    "from": "gpt",
    "value": "My name is Vicuna, and I'm a language model developed by Large L"
  }
]
```

Performance



How?

Follow Vicuna

- Dataset
- LLaMA model parameters
- Delta of Vicuna parameters (optional)
- Fine-tuning

Datasets

```
{  
  "id": "identity_1",  
  "conversations": [  
    {  
      "from": "human",  
      "value": "Lab meeting on 5/12/2023"  
    },  
    {  
      "from": "gpt",  
      "value": "Recorder: yuzhi\n  Next week - Summary goals for summer\n  Cor"    }  
  ]  
}
```

LLaMA and Vicuna model parameters

- Follow https://huggingface.co/docs/transformers/main/model_doc/llama
- `/data/public/LLaMA/download_community.sh`
 - run with `./download_community.sh 7B /save/path`
- `/data/public/LLaMA/*`

Fine-tuning

Follow: <https://github.com/lm-sys/FastChat>

Run with:

```
torchrun --nproc_per_node=4 --master_port=20001 fastchat/train/train_mem.py \
  --model_name_or_path ~/model_weights/llama-7b \
  --data_path playground/data/dummy.json \
  --bf16 True \
  --output_dir output \
  --num_train_epochs 3 \
  --per_device_train_batch_size 2 \
  --per_device_eval_batch_size 2 \
  --gradient_accumulation_steps 16 \
  --evaluation_strategy "no" \
  --save_strategy "steps" \
  --save_steps 1200 \
  --save_total_limit 10 \
  --learning_rate 2e-5 \
  --weight_decay 0. \
  --warmup_ratio 0.03 \
```


Problem

- The size of tensor a (65537024) must match the size of tensor b (262148096)
 - Need more data to fit batches
- RuntimeError: CUDA out of memory
 - change
`python3.10/site-packages/torch/distributed/fsdp/_state_dict_utils.py`
 - `state_dict[fqn].clone().detach()` to `state_dict[fqn].cpu().clone().detach()`

Results

Ask Who is the Presenter on 4/14/2023 lab meeting?

- Playgroud data + Lab notes
- Lab notes only

4/14/2023

Presenter: Yuwei

Recorder: Nasy

Topic: Reinforcement Learning from Human Feedback

1. Basic Reinforcement learning
 - a. On-policy
 - i. What is policy? Examples.
 - b. Off-policy
2. RL without well-specified reward function
3. RLHF and PPO (Proximal policy optimization)

Results

- Playgroud data + Lab notes with 3 Epochs

```
➤ CUDA_VISIBLE_DEVICES=2 nixGL pdm run python -m fastchat.serve.cli --model-path outputs/tr8 --style rich
Loading checkpoint shards: 100%
```

Human:

Who is the Presenter 4/14/2023 lab meeting?

Assistant:

Wenliang Zhong

Human:

Results

- Playground data + Lab notes with 50 Epochs

```
/data/public/FastChat on main !1 ?7 .....  
CUDA_VISIBLE_DEVICES=2 nixGL pdm run python -m fastchat.serve.cli --model-  
Loading checkpoint shards: 100%|
```

Human:

Who is the Presenter 4/14/2023 lab meeting?

Assistant:

I'm a language model called SMILE-lab-ai, and I'll be presenting some research

Human:

Results

- Lab notes with 3 Epochs

```
~/data/public/fastchat on ☪ main:1.7
nixGL pdm run python -m fastchat.serve.cli --model-path outputs/tr7 --style rich
Loading checkpoint shards: 100%
```

Human:

Who is the Presenter 4/14/2023 lab meeting?

Assistant:

The presenter for the 4/14/2023 lab meeting is Yuwei Miao.

Human :

Results

- Lab notes with 50 Epochs

```
nixGL pdm run python -m fastchat.serve.cli --model-path outputs/tr5 --  
Loading checkpoint shards: 100%
```

Human:

Who is the Presenter 4/14/2023 lab meeting?

Assistant:

- 1 Presenter: Qifeng Zhou
- 2 Recorder: Wenliang Zhong
- 3 Topic: GPT-2 Implementation Details
- 4 Questions
 - 1 How to calculate Q, K, V? When are they different?
 - 2 How to acquire bbpe encoding?
 - 3 Where are the multi heads? How to calculate heads?
 - 4 BBPE how to separate words?

Human:

Conclusion

- Which model to choice?
 - From the rank <https://lmsys.org/blog/2023-05-10-leaderboard/>
- How to fine-tune?
 - Follow Vicuna FastChat framework
 - Follow Alpaca instruct framework

Reference I

- [1] Noam Shazeer. *GLU Variants Improve Transformer*. Feb. 12, 2020. DOI: [10.48550/arXiv.2002.05202](https://doi.org/10.48550/arXiv.2002.05202). arXiv: [2002.05202](https://arxiv.org/abs/2002.05202) [cs, stat]. URL: <http://arxiv.org/abs/2002.05202> (visited on 05/19/2023). preprint.
- [2] Jianlin Su et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. Aug. 8, 2022. DOI: [10.48550/arXiv.2104.09864](https://doi.org/10.48550/arXiv.2104.09864). arXiv: [2104.09864](https://arxiv.org/abs/2104.09864) [cs]. URL: <http://arxiv.org/abs/2104.09864> (visited on 05/19/2023). preprint.
- [3] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. Feb. 27, 2023. arXiv: [2302.13971](https://arxiv.org/abs/2302.13971) [cs]. URL: <http://arxiv.org/abs/2302.13971> (visited on 05/19/2023). preprint.

Example

Outside.