Y **Hacker News** new | threads | past | comments | ask | show | jobs | submit kordlessagain (2479) | logout

A simple guide to fine-tuning Llama 2 (brev.dev)

280 points by samlhuillier 27 days ago | hide | past | favorite | 53 comments

nmitchko 27 days ago | next [–]

This is a pretty useless post. You could also follow the same 1000x tutorials about llama and use the already uploaded hugging face formats that are on hugging face...

Here are some actually useful links

https://blog.ovhcloud.com/fine-tuning-llama-2-models-using-a...

https://huggingface.co/meta-llama/Llama-2-70b-hf

https://huggingface.co/meta-llama/Llama-2-7b-hf

> onlypositive 27 days ago | parent | next [–]
>
> Is it really "useless" if I didn't even know about llama? And look, now I have 3 more links to dive into.
>
> This is the opposite of useless.
>
>> mciancia 27 days ago | root | parent | next [–]
>>
>> Well it's quite possible that it is useless for you since you didn't hear about llama by now ;)
>>
>>> hoten 27 days ago | root | parent | next [–]
>>>
>>> I guess this was meant as a tongue-in-cheek joke comment, but it comes off as needless gatekeeping.
>>>
>>>> onlypositive 26 days ago | root | parent | next [–]
>>>>
>>>> I dunno what's worse, the pointless commentary, needless gatekeeping, the superfluous white knighting or the fact we're getting upvotes for all this nonsense.
>>
>> DigiDigiorno 27 days ago | root | parent | prev | next [–]
>>
>> Three links is not a large number of links, AND you don't "have" to dive into every link in lists of "useful links".
>>
>> That said, I don't think op is useless.
>
> alfalfasprout 27 days ago | parent | prev | next [–]
>
> Honestly HN has turned into reposting random how-tos about the latest hype LLMs.
>
>> g42gregory 27 days ago | root | parent | next [–]
>>
>> If that's what people like, why is that a concern?
>>
>> hamilyon2 27 days ago | root | parent | prev | next [–]
>>
>> HN hyped machine learning long before the current craze cycle
>
> IAmNotACellist 27 days ago | parent | prev | next [–]
>
> And it's probably already in TextGen and better UIs for serious stuff than yet another Jupyter notebook.
>
> 3abiton 27 days ago | parent | prev | next [–]
>
> I still don't get why it's useless. Maybe less useful, but useless seems to be a hyperbole.
>
>> nmitchko 27 days ago | root | parent | next [–]
>>
>> For each step in the article:
>>
>> 1. You should really download the huggingface hosted model.
>>
>> 2. Why convert the model if meta already hosts it here: https://huggingface.co/meta-llama
>>
>> 3. This step completely glossed over the hardware requirements. Doesn't explain any of the instructions needed in the finetuning process.
>>
>> 4. "Now run the model". What? What about the model precision, hosting, standard open source tools, etc,etc...
>
> jeremycarter 27 days ago | parent | prev | next [–]
>
> Thanks!

zoogeny 27 days ago | prev | next [–]

What I'd like to do is create a website where:

1. There is a list of open source fine-tuning datasets on millions of topics. Like, anime, lord of the rings, dnd, customer service responses, finance, code in many programming languages, children's books, religions, philosophies, etc. I mean, on every topic imaginable sort of like a Wikipedia or Reddit of fine-tuning data sets.

2. Users can select one or more available datasets as well as upload their own private datasets

3. Users can turn-key fine-tune llama 2 or other pre-trained models

Right now, doing this kind of thing is way beyond the capability of the common user.

IAmNotACellist 27 days ago | parent | next [–]

I personally don't see a future where common users will ever have to know the phrase "fine-tuning" or worry about it. The most I can see is "Do you consent to share your information with Apple/Meta/X/Microsoft/OpenAI's knowledge engine?" and if you agree, everything they have on you will power an extremely powerful all-encompassing knowledge engine. Probably with some daily recommendations to integrate a new domain into it, like, "We noticed you're into Lord of the Rings, so we went ahead and made your knowledge engine familiar with the collected works of Tolkein, all historical academic and modern interpretations and criticisms, transcripts of the movies, and generative AI fan fiction capabilities."

zoogeny 27 days ago | root | parent | next [–]

I don't think the major barrier to the idea would be consumer awareness. For the near-term the major barrier will be cost. Just as one example, together.ai offers fine-tuning service at an advertised cost of $0.001 per 1k tokens used [1]. That will get pricey for even small datasets. No doubt this will come down, but I don't see consumers paying $1000 for a customized AI model that they then have to pay inference costs to run. Maybe once we get consumer devices that have sufficiently capable AI accelerators (e.g. Apple Neural Engine) to run sufficiently capable llm models, then customers would be willing to customize and run local.

The second point is, we don't know if fine-tuned models, vector search or more-massive general purpose llm models is the right way to go.

But for business-to-business, I think this might be a viable business. If you had a whole bunch of ready-to-go open-source fine-tune datasets for commercial applications you might find a market of businesses that want to run their own models for a variety of reasons.

1. https://together.ai/pricing

samstave 27 days ago | root | parent | prev | next [–]

The year is 2149, previously, we thought *time* was the real commoditiy, *water* before that, and *money* even prior....

But now. Now. Its DNAX... Cloning fraudulant DNA to make BIO-chips to unlock credits for "yee ol' goods an' services gub'nah"

Basically every transaction is bio-tracked, so if you want an off grid you have to have false clones...

DNA from old embryoes that allow you to build identities in their names and wear them like sleaves to navigate the systems.

This is how you manipulate the engines.

brucethemoose2 27 days ago | parent | prev | next [–]

This sounds like a great fit for Cerebras, if they can set up the text database front end.

They could host the text database for free, and then offer a "oh look, you can train llama on this text *right now* for cheaper than a Nvidia box" button on every listing.

Then charge through the nose for private business training (kinds like they do now, but charging more.)

zoogeny 27 days ago | root | parent | next [–]

I agree that it would be almost impossible to defend this kind of business, especially if you stayed committed to open-source datasets. It would come down to the UX and the community if you hoped to survive. Probably long-term you would either have to get into your own pre-trained models, fight the commodity hosting business or aim to get acquired.

brucethemoose2 27 days ago | root | parent | next [–]

Well civitai is basically what you are describing. Its very doable.

But a big difference is that (for now) Stable Diffusion finetuning is much easier than LLaMA.

samstave 27 days ago | parent | prev | next [–]

ELI5 - who exactly makes the open datasets you refer to? [SERIOUS Q]

zoogeny 26 days ago | root | parent | next [–]

This would initially be a community, like Wikipedia, Reddit, Github, etc. People who are passionate about the future of AI, believe in the value of open source data and want their voice to be part of a community of data that will be used to train AIs in the future.

In my wildest dreams, and even reasonably, you could incentivize people with a digital currency. I was thinking something along the lines of a community that could stake some money ($100/$1000). They would then get "ownership" and moderating rights to the contents of a dataset. Other people could submit content to their dataset that they could allow or deny. The allowance of the content would distribute some share of the stake in the form of tokens. Then they would be able to re-sell the data in the set to people who want to fine-tune AIs using that dataset. The value of the tokens associated with that dataset would go up thereby distributing some portion of the profit to the moderators and the contributors.

syntaxing 27 days ago | prev | next [–]

Can someone share a good tutorial how to prepare the data? And for fine tuning, does a 3090 have enough VRAM? I want to do what the author mentioned by fine tuning the model on my personal data but I’m not sure how to prepare the data. I tried using vector search + LLM but I find the results very subpar when using a local LLM.

jawerty 27 days ago | parent | next [–]

I just streamed this last night https://m.youtube.com/watch?v=TYgtG2Th6fI&t=3998s

I’ve been live streaming myself fine tuning llama on my GitHub data (to code like me)

jeremycarter 27 days ago | root | parent | next [–]

Fantastic job! Very easy to follow

jawerty 27 days ago | root | parent | next [–]

Thank you! I have some other streams where I do little projects like these check them out

notpublic 27 days ago | parent | prev | next [–]

As mentioned in the OP's blog post, checkout https://github.com/facebookresearch/llama-recipes.git. specifically files in ft_datasets directory.

I am able to finetune meta-llama/Llama-2-13b-chat-hf on a 3090 using instructions from quickstart.ipynb.

syntaxing 27 days ago | root | parent | next [–]

Oh interesting, I didn’t know the documentation expanded so much in this past month.

samlhuillier 27 days ago | parent | prev | next [–]

Working on this now!

syntaxing 27 days ago | root | parent | next [–]

I’m looking forward to this! Are you using an adapter (I don’t see it mentioned in your article)? I was under the impression you cannot fit 7B at 4 bit since it’ll take 25GB of VRAM or so.

samlhuillier 27 days ago | root | parent | next [–]

Yes using the qlora adapter that hugging face provides with peft

syntaxing 27 days ago | root | parent | next [–]

ahh, I was on my phone before so I must of glimpsed over it, I see it on the last section. Thanks!

marcopicentini 27 days ago | prev | next [–]

Anyone has calculate the break even point (as number of token per month) between self-hosted LLAMA and OpenAI GPT-3.5 API?

eachro 27 days ago | prev | next [–]

I've veen a bit out of the loop on this area but would like to get back into it given how much has changed in the LLM landscape in the last 1-2 yrs. What models are small enough to play with on Collab? Or am I going to have to spin up my own gpu box on aws to be able to mess around with these models?

naderkhalil 27 days ago | parent | next [–]

Hey, you could use a template on brev.dev to spin up a gpu box with the model and Jupyter notebook. Alternatively, the falcon 7b model should be small enough for colab

treprinum 27 days ago | prev | next [–]

Is there any tutorial on how to use HuggingFace LLaMA 2-derived models? They don't have checkpoint files of the original LLaMA and can't be used by the Meta's provided inference code, instead they use .bin files. I am only interested in Python code so no llama.cpp.

lolinder 27 days ago | parent | next [–]

I'd reconsider your rejection of llama.cpp if I were you. You can always call out to it from Python, but llama.cpp is by far the most active project in this space, and they've gotten the UX to the point where it's extremely simple to use.

This user on HuggingFace has all the models ready to go in GGML format and quantized at various sizes, which saves a lot of bandwidth:

https://huggingface.co/TheBloke

treprinum 27 days ago | root | parent | next [–]

I understand, I use llama.cpp for my own personal stuff but can't override the policy on the project I want to plug it in, which is python-only.

pests 27 days ago | root | parent | next [–]

There was a post yesterday about a 500 line single-file C implmenetation of llama2 with no dependencies. The llama2 architecture is hard coded. It shouldn't be too hard to port to python.

Found the repo, couldn't easily find the HN thread.

https://github.com/karpathy/llama2.c

treprinum 27 days ago | root | parent | next [–]

That's Andrej's toy project, it won't run 7B LLaMA.

NateGenX 25 days ago | root | parent | next [–]

It does now: https://github.com/karpathy/llama2.c#metas-llama-2-models

ramesh31 27 days ago | parent | prev | next [–]

>I am only interested in Python code so no llama.cpp.

llama cpp has python bindings: https://pypi.org/project/llama-cpp-python/

Here's using it with langchain: https://python.langchain.com/docs/integrations/llms/llamacpp

kiratp 27 days ago | parent | prev | next [–]

If you want high performance inference us this: https://vllm.ai/

For local batch stuff: https://huggingface.co/docs/transformers/main_classes/pipeli...

moneywoes 27 days ago | prev | next [–]

Any fine tuning success stories? Or real world use cases

bvm 27 days ago | parent | next [–]

Sure. I worked at a company that produced tens of thousands of human written summaries of news data a year. This was costly and slow but our clients really valued them. Back in 2019 we fine tuned an LLM to help, we put a lot of effort into creating a human-in-the-loop experience, highlighting parts of speech that were commonly hallucinated and ensuring that we were allowing humans to focus on things that humans are good at.

We also released some of the data as a free dataset with a commercial option for all of it. This was more successful than I thought it would be and was hoovered up by the kind of people that buy these datasets.

It will have been surpassed by recent developments now but it was an incredibly enjoyable project.

rmbyrro 27 days ago | root | parent | next [–]

What kind of clients value news summaries that much?

bvm 27 days ago | root | parent | next [–]

large corporates, financial services. Use cases were needle-in-a-haystack style searching, internal comms, following research topics over time, external newsletters, that kinda stuff. It wasn't particularly high margin but it was a fun business.

mothcamp 27 days ago | root | parent | next [–]

I imagine it's a company similar to Bulletin Intelligence. Would you be open to discussing your experiences in this industry?

bvm 27 days ago | root | parent | next [–]

yeh sure! how'd you like to do so?

mothcamp 27 days ago | root | parent | next [–]

Awesome! To protect your privacy on HN, please email nparker2050@gmail.com and let me know whether you prefer getting on a call or keep things in writing. Looking forward to hearing from you!

m00dy 27 days ago | prev [–]

Which dataset would be good to fine-tune for developing sales assistant like chatbot ?

ShamelessC 27 days ago | parent [–]

You could try using a transcript of The Wolf of Wall Street, maybe throw in Glengarry Glen Ross for good measure?

/s

CamperBob2 27 days ago | root | parent [–]

First prize is an 80 GB H100. Second prize is a 4090. Third prize is a PIP.