OpenLLM Talk 002

序章

背景介绍

【缘起】：OpenLLM Talk 这个事情起源于 20230603 OpenLLM 交流群中大家的一场讨论，本着心动不如行动的想法，我们花了一点时间来将其落地，希望可以为大家提供一个 LLM/NLP 领域的**交流平台**。——**我们或许不够 AI，但尽量足够 Open；我们也不知道能走多远，但尽量比自己想的更远。**

【结构】：整体上分成本周新闻、本周推荐、本周经典（可选）、本周实践（可选）、free talk 等版块，建议后续最好采用每个版块每期由 1-2 人认领+多人参与贡献+自由讨论的形式。

本期记录

【编号】：OpenLLM Talk 002 (三位数是希望 LLM 的热度+我们的热情+读者的热情可以支撑我们做到三位数)

【时间】：20230701 晚上九点（每周六晚上九点，节假日顺延）

【本期提要】：chatgpt 增速放缓；llm 用作命令行；RoPE 的上下文窗口拓展；vllm vs llama.cpp；lora 融合；参数和数据比例；

【本期贡献者】- 排名不分先后：

【主持人】：羡鱼（后续每期由大家自行认领）

【版块负责人】：羡鱼、yuhan、suc16 （后续每期由大家自行认领）

【具体内容贡献者】：请查看具体内容后面的署名，比如问题、回答和观点的来源

【talk 视频】：后续放出

注意事项

【talk 模板】：https://zhuanlan.zhihu.com/p/640522290；可参考模板进行贡献

【小要求】：版块负责人认领之后尽量准时参加，其余同学可自行选择是否参与；

本周新闻

【本周新闻】：LLM/AI news，包括但不限于学术、项目、工业界新闻和进展；多人认领或者直接在此添加，由 **1-2 人认领并汇总**；建议大家都参与进来，相互补充，尽量**减少信息冗余和缺漏**；共~10 分钟；

【贡献者】：yuhan、羡鱼

【建议区】：可以考虑 GitHub 的讨论区，看个人习惯；论文可以写个摘要；

学术

项目

1.gorilla-cli

简介：

https://github.com/gorilla-llm/gorilla-cli

使用 llm 作为命令行助手

Usage

Activate Gorilla CLI with a straightforward `gorilla` followed by your command in plain English.

For instance, to list all files in the current directory, type:

```
$ gorilla I want to list all files in the current directory
```

or if you prefer, you can use quotes to avoid issues with string parsing:

```
$ gorilla "I want to list all files in the current directory"
```

Gorilla CLI will then generate potential commands. Simply use the arrow keys to navigate through the options, then press enter to execute the chosen command.

```
Welcome to Gorilla. Use arrows to select
» ls
  ls -l
  ls -al
```

工业界

ChatGPT 访问量增速大降，6 月环比增长率可能为负数，科技股资金大幅外流，AI 热潮熄火了吗？ - 知乎

https://www.zhihu.com/question/608894843

ChatGPT 访问量增速大降；明星 L4 卡车被曝停摆：清华编程天才合伙，创办仅 19 个月；B 站决定取消播放量显示 | 雷峰早报 - 雷峰网的文章 - 知乎

https://zhuanlan.zhihu.com/p/639930168

本周推荐

【本周推荐】：本周重点内容推荐和介绍，模型、开源项目、好的资料或课程，建议 1-3 项；共 15 分钟；

【贡献者】：suc16

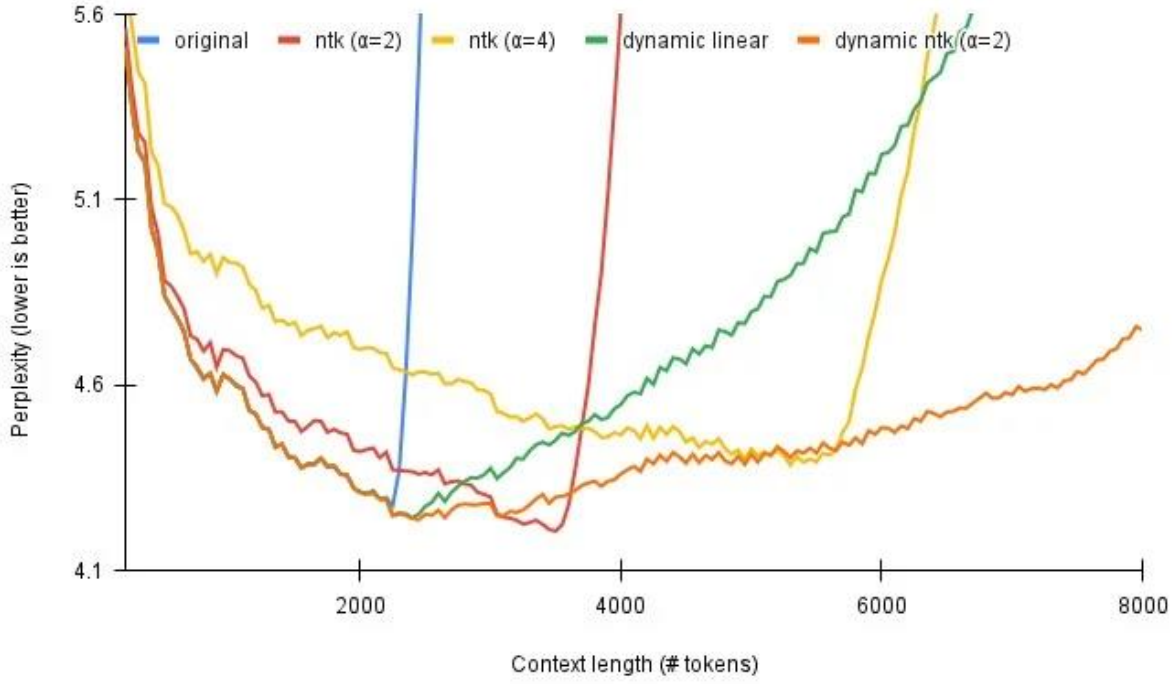
【提名区】：

【建议区】：

【本期主题】：RoPE 的上下文窗口拓展

RoPE 的上下文窗口拓展，**从线性位置插值到非线性位置插值；**

https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/



模型

项目

杂项

示例：可参考 OpenLLM Talk 001 中的 state of gpt

refs:

https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/

本周经典-optional

【本周经典】：NLP/LLM 领域的经典话题探讨；~15 分钟；

【贡献者】：

【提名区】：位置编码、量化

【本周主题】：

本周实践-optional

【本周实践】：NLP/LLM 领域实践经验分享，可以分享自己的实践经验或者他人的实践经验，后面群里也会组织一些实践内容；~15 分钟；

【贡献者】：yuhan

【提名区】：

【建议区】：coding 搞起来；后续拉个 read_code/paper 分支，LLM 精读、注释；专门建一个**数据专题**；

1.vllm vs llama.cpp

结论：**vllm 速度是 llama.cpp 16fp 速度的 3 倍**，llama.cpp 不支持批处理，vllm **支持批处理**；

单张 v100 vllm 13B 推理：16-18token/s，对输入长度不敏感；--yuhan

batch size：v100 可以开到 32

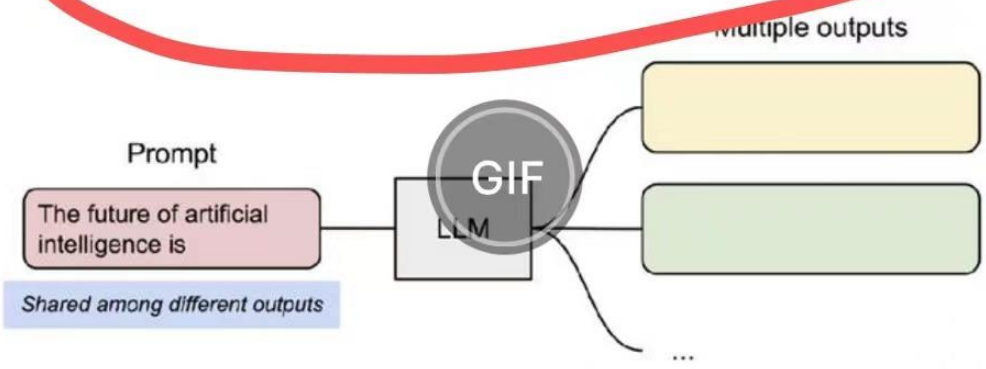
vLLM on V100　　64->64:19 词每秒 bz==1　　2048-256:16 词每秒 bz==1 （前缀相同可以复用提速）
batch size == 32 时，最快可达 300+词每秒
当输入 token 为 2048 时，32batch 和 1batch 速率一直，在此之前，tps 随 batch 数线性增长
（不过上述实验存在很大缺陷：每个 batch 使用的 prompt 相同，后续会对实验进行调整）

vLLM 目前存在不支持 context 超过 4096 的情况，后续解决这一系列问题的话，是一个非常值得尝试的推理框架

而这种内存效率的提升，能让系统将更多的序列进行批处理，提高GPU利用率，从而显著提高吞吐量。

此外，PagedAttention还具有另一个关键优势：高效的内存共享。

比如在并行采样中，就能从相同的提示生成多个输出序列。在这种情况下，提示的计算和内存可以在输出序列之间共享。



并行采样的示例

PagedAttention通过块表自然地实现了内存共享。

类似于进程共享物理页的方式，

Free Talk

【Free Talk】自由提问，自由讨论；在文档里提问或者在群里提问，建议尽量在此汇总；如果群里已经有比较好的讨论结果，也可以将讨论结果搬运过来；时间不限；

【贡献者】：羡鱼（编辑）+OpenLLM 群友

线上讨论:

Q1：目前有哪些比较好用的推理框架？

A1：MLC/lamma.cpp/fastllm/vllm

Q2：RoPE 的外推性？

A2：RoPE 的长度外推性并不好；具体的外推方案后续可以再具体讨论；

Q3：deepspeed-chat 训练 7B 的最小资源需求？

A3：RLHF 可能很多人都还没搞明白；

RM 应该更大吗？好像之前也听到过 RM 可以小一点儿？

智源大会上有人说 RM 应该比较强才会有好的效果--刘鹏飞

目前对 RLHF 比较了解的：OpenAI、Anthropic 的 claude 模型、deepmind

Q4: chatglm2 是 decoder-only 吗？

A4：暂不清楚；

GLM：以自回归的方式来做完形填空

GLM2 项目说之后会发 paper。

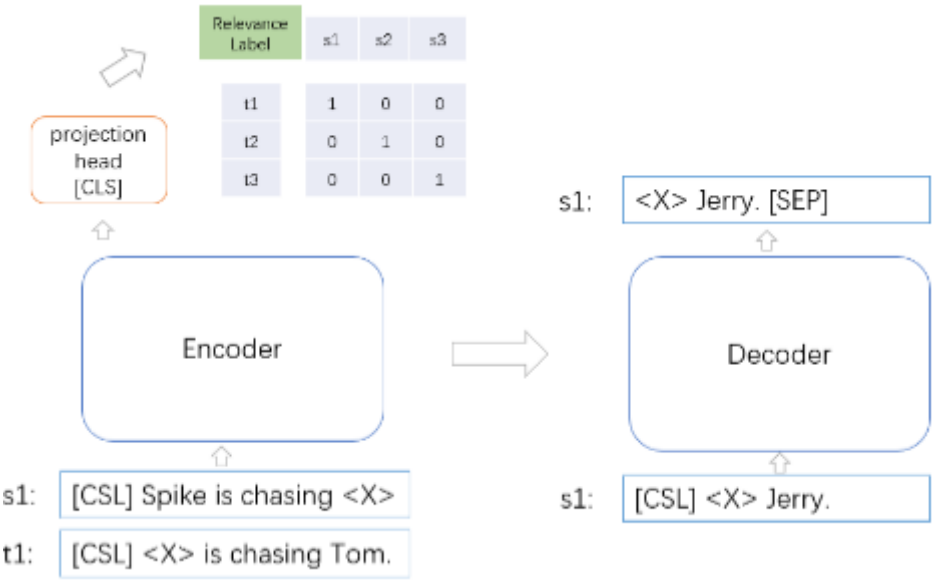
Q5：大家在语言链中检索时使用哪些向量表示模型？

A5：

GUR：Generate to Understand for Representations--2023

预训练时带一个 SE 任务，SE 任务用的对比学习；

https://arxiv.org/abs/2306.10056



直播分享 7/6 19:00-20:00
点击链接入群收看：
https://work.weixin.qq.com/gm/aa9a0a500d650aee6f99ce3d6f503cfb

Q6: 模型参数和训练语料的合适比例？

A6：OpenAI 1:1.7　gpt3？

　　Deepmind 1:20

这两个数据的来源可以把链接放上来吗--来自何枝大佬的分享

数据和模型的硬盘体积比，无论什么语言模型：10-100 倍？在小模型到 BERT 这个量级上做的；按这个预估 1T 数据，大约 100G 左右的模型（50B）；

GPT3 的数据：大约 400B token，570G，参数 175B（175B model, 350B data?）；可能数据量相对模型来说有点儿小；

大规模中文数据：

wudao；

Q7：大模型如何控制表示的精确性？减少幻觉

A7：https://export.arxiv.org/pdf/2306.03341v2.pdf

检索增强；苏神 NBCE 里提到过一篇文章；

推理时改 head 权重；Inference-Time Intervention - 林知的文章 - 知乎

https://zhuanlan.zhihu.com/p/637317327

任务拆分，用工具；比如数学四则运算能力，不如调用工具--符尧

lets verify step by step

Step-wise RLHF?

Q8：LLM 训练不充分是为什么？模型太大还是数据不够？

A8：主要还是数据不够；

Train Large, Then Compress

https://arxiv.org/abs/2002.11794

Q9：不太卷的 LLM 组？

A9：外企和国企：NV 少量 HC；国企，XX 保险；

Q10：LLM 基座及训练选型？

A10：

同量级 llama 强于 bloom；

Openbuddy 项目目前使用的基座是 llama 和 falcon

训练：

trl

deepspeed

lora：PEFT，lora 比全参数微调差；lora 的学习率需要更高？lora 的灾难性遗忘问题不那么严重？

lora 秩的选择，为啥都比较小，经验选择，开大了没啥用；图像领域的 lora rank 可能会比较大，比如 128、256 等等。

上次讨论过的问题：

NLP 的 lora 融合，已经有一些工作了：

https://arxiv.org/abs/2306.14870

群里讨论：

有空再同步；

参考资料

https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/

https://arxiv.org/abs/2306.10056

https://github.com/gorilla-llm/gorilla-cli

NLP 的 lora 融合：

https://arxiv.org/abs/2306.14870

ChatGPT 访问量增速大降，6 月环比增长率可能为负数，科技股资金大幅外流，AI 热潮熄火了吗？ - 知乎

https://www.zhihu.com/question/608894843

ChatGPT 访问量增速大降；明星 L4 卡车被曝停摆：清华编程天才合伙，创办仅 19 个月；B 站决定取消播放量显示 | 雷峰早报 - 雷峰网的文章 - 知乎

https://zhuanlan.zhihu.com/p/639930168

后续计划

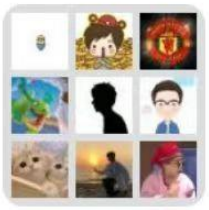
- 正式开启 OpenLLM talk 的运营，P1；
- ChatPiXiu 项目：陆续有一些实践计划，P0；

- OpenSE：检索项目，P0，字符检索+语义检索；
- OpenLLM：LLM 学习和实践项目
- OpenAIDic：科普项目

加入/赞助我们

我们非常缺人，也非常缺时间和算力，希望能有越来越多的朋友参与进来，认领 talk 的组织者、主持人（最近从杭州跑北京来了，工作比之前忙不少，不太可能每期都由我来组织了~）、版块的负责人；参与项目后续的开发和讨论等等。

微信群：（请优先加入微信群，如果失效则加入 QQ 群再私聊我进微信群）



群聊：羡鱼智能-OpenLLM技术交流群

该二维码7天内(7月7日前)有效，重新进入将更新

QQ 群：

羡鱼智能-OpenLL...

群号：740679327

扫一扫二维码，加入群聊。

QQ