OpenLLM Talk 007

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 007 (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】：20230805 晚上九点（一般每周六晚上九点，节假日顺延）

【本期提要】：DeepSpeed-Chat；中文 llama2；通义千问开源；模型部署；数学能力；OpenLLMAI

【本期贡献者】- 排名不分先后：

【主持人】：（后续每期由大家自行认领）

【编辑】：（最好由主持人兼任）

【版块负责人】： （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】：

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】：主持人及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】：

【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

注：论文+重点

DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales

https://arxiv.org/abs/2308.01320

项目

YuLan-Chat-2：基于 LLaMA-2 的全新中英文对话大模型

https://mp.weixin.qq.com/s/VciHxcQEdJ9APBNBK4xk4g

工业

中文 llama2

https://github.com/ymcui/Chinese-LLaMA-Alpaca-2

免费、可商用，阿里云开源 70 亿参数通义千问大模型

https://mp.weixin.qq.com/s/hzAxiTeiJlLGzDDB0kSOxg

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】:

【提名区】:

【建议区】:

【本期主题】:

资料

LLM-QAT
https://github.com/facebookresearch/LLM-QAT

权重量化对大模型涌现能力的影响
https://zhuanlan.zhihu.com/p/647347411

【LLM 003】 并行训练汇总 - JOYWIN 的文章 - 知乎
https://zhuanlan.zhihu.com/p/647133493

模型

MSRA 提出新架构 RetNet，将取代 Transformer
https://www.xiaohongshu.com/discovery/item/64c5f19800000000017019463?app_platform=android&app_version=7.96.1&author_share=2&share_from_user_hidden=true&type=normal&xhsshare=WeixinSession&appuid=61eab616000000000201974a&apptime=1690695042

项目

https://github.com/huggingface/transformers/pull/24653
这个可以看，现在最新的版本是 dynamic NTK 了，现在已经集成到最新的 huggingface 里的 llama

杂项

refs:

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】：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】：

【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**；

OpenLLMAI 实践计划与组织分工

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:

1. llama-2-70b 部署问题。

   使用 a800-80g 卡部署 8bit 量化后，大约一秒一个 token，巨慢，从这个角度说，chatgpt 这种 175b 能这么快是真的很神奇。希望大家讨论一下

答：两张卡，GPU 占用间歇性能到 100%，pipeline 推理；可以换 deepspeed 试试；

2. 精调版本问题，现在除了更换基底模型以为，实际上不同的精调策略出来的模型能力非常多样性。例如 llama-70b 有非常多的精调版本，mt bench 从 6.4 到 7.4 不等，部分超越官方的 chat，小模型的精调版本就更多了。除了更进一步的 rlhf，有什么关键性的共识可以帮助大家提升性能？顺带一提，vicuna 都 1.5 了。

答：以 MMLU 为例，不同的精调模型效果为何差距很大？精调有什么技巧？

3. 对齐税？

答：

对齐税/负对齐税？

防止灾难性遗忘，比如 RLHF 阶段带上预训练数据和目标；

领域数据和通用数据的混合比例，可以参考链家的实验 chathome；

4. 预训练 long context？

答：

塞满 token 限制，为了训练效率；

5. llama 安全限制？

答：llm attack

6. GPT3.5 变快了？chatgpt 网页和 API 结果不太一样？

答：压缩、蒸馏？

或许不是一个模型

https://mp.weixin.qq.com/s/NFhacKQRG7_7ltQQO_pnkw

7. 数学能力：GPT4 很强 4-5/5，llama2 1-2/5，GPT3.5 0、5?

答：GPT3.5 可能并不够强

8. llama2 的 license 问题？

答：

9. RLHF 可能会影响 SFT 或者基座的效果？

10.

群里讨论：

有空会同步，取决于人力，希望大家积极认领~

参考资料

后续计划

我们正式升级为一个不太正式的组织了！叫做 OpenLLMAI.

https://github.com/OpenLLMAI

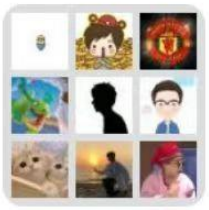
- 正式开启 OpenLLM talk 的运营，P1；
- ChatPiXiu 项目：陆续有一些实践计划，P0；
- https://github.com/OpenLLMAI/OpenLLaMA2，P0，doing
- https://github.com/OpenLLMAI/chinese-llama2，P0，doing
- OpenSE：检索项目，字符检索+语义检索，P0；
- OpenLLM：LLM 学习和实践项目，P0；
- OpenAIWiki：AI wiki for everyone；
- ChatLover：模拟恋人+爱情助手，P1；

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）

（二维码过期了 !）



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

往期精彩

【OpenLLM Talk 006】本期提要：LLM 加水印；softmax 的 bug；llama2 汉化；多轮对话；DPO 论文阅读；LLM 评估；SE；量化；NOPE；长度外推；OpenLLMAI 与实践计划 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/647879679

【OpenLLM Talk 005】本期提要：llama2；FreeWilly；LLM 推理与评估；LLM 八股；RetNet；DPO；数据配比 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/645679737

【OpenLLM Talk 004】本期提要：外挂知识；抱抱脸每日论文；MOSS-RLHF；GPT4 细节；OpenAI 代码解释器；百川 13B；LLM 面经；多轮对话；数学能力；反思；LLM 中的知识 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/643960837

【OpenLLM Talk 003】本期提要：SuperCLUE-Open；文心盘古；chatlaw；LLM 综述；NTK-Aware Scaled RoPE；10 亿上下文；InternLM；GLM 讲座 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/642376781

【【OpenLLM Talk 003】SuperCLUE-Open；文心盘古；chatlaw；LLM 综述；NTK-Aware Scaled RoPE；GLM 讲座】 【精准空降到 10:10】
https://www.bilibili.com/video/BV1Kh4y1E7nX/?share_source=copy_web&vd_source=9e7882f0ef2735e23d66a6f128612943&t=610

【OpenLLM Talk 002】本期提要：chatgpt 增速放缓；gorilla-cli；RoPE 外推；vllm vs llama.cpp；lora 融合；模型参数和数据之比；OpenSE 计划 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/641285737

【OpenLLM Talk 001】本期提要：长程记忆；OpenAI 上新；百川智能 7B 模型；State of GPT；位置编码；deepspeed-rlhf；RLHF 数据 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/640275116

【OpenLLM Talk 000】我们做了一个 LLM 领域的交流平台 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/636350755

【OpenLLM Talk 模版】兴趣和热爱胜过一切，OpenLLM 就从这里开始吧！欢迎加入！ - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/640522290