OpenLLM Talk 012

注意事项

出于隐私保护和数据安全的考量，建议尽量不要在 talk 过程中涉及到自己的单位信息、自己的隐私信息、违反法律和道德的信息以及其他引起争议的内容，请保护好自己的马甲。

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 012 (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】：20230909 晚上八点（一般每周六晚上八点，节假日顺延）

【本期提要】：LLM 压缩综述；高效 RLHF；OpenPPL；百川 2；bytepiece；开发者日志；组织建设

【本期贡献者】- 排名不分先后：

【主持人】：Hope、羡鱼（后续每期由大家自行认领）

【编辑】：羡鱼（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

大模型压缩首篇综述来啦~

https://zhuanlan.zhihu.com/p/652434165

tensorrt llm 支持很多量化算法，也支持并发

基于大语言模型的自主智体

https://zhuanlan.zhihu.com/p/654609159

Efficient RLHF: Reducing the Memory Usage of PPO

https://huggingface.co/papers/2309.00754

项目

OpenPPL-LLM | OpenPPL 之大语言模型推理引擎来啦

https://zhuanlan.zhihu.com/p/653808774

工业

LLaMA 核心原作多半离职，Meta AI 内幕曝光！算力争夺撕破脸，大模型团队成员连换三轮

https://zhuanlan.zhihu.com/p/654619099

09 月 06 日 Baichuan2 发布，开源 7B 和 13B 模型，使用体验如何，将给行业带来那些影响？

https://www.zhihu.com/question/620743448

中间 checkpoint 有什么作用

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

模型

项目

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】：dev 层的内容以后都放到实践部分；

OpenLLMAI 开发者日志：

【OpenLLM Dev007】当前进展及开发计划-SEP01

框架层：

https://github.com/OpenLLMAI/OpenLLaMA2

已经跑通 llama2 7B 全流程；

data 层：

https://github.com/OpenLLMAI/OpenLLMData

建立了初步的工作流；

模型层：

产出了初版 toy model，基本复刻了 Chinese-alpaca2 在 ceval 上的的效果

https://cevalbenchmark.com/static/model.html?method=MoYu

这名字可能有点儿咸鱼了，以后管它叫**星海**？

57	MoYu	OpenLLMAI	2023/8/27	40.3

其他：

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

https://arxiv.org/abs/2306.14048

线上讨论:

1. baichuan2 放出的中间 checkpoint 可以做什么
2. https://spaces.ac.cn/archives/9752/comment-page-1

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 系列的运营，P1；
- ChatPiXiu 项目：陆续有一些实践计划，现已分拆为各个项目，貔貅只做文档，P1；
- https://github.com/OpenLLMAI/OpenLLaMA2，P0，doing
- https://github.com/OpenLLMAI/chinese-llama2，P0，doing
- https://github.com/OpenLLMAI/OpenLLMData，P0，doing
- OpenSE：检索项目，字符检索+语义检索，P1；
- 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；

组织建设

加入/赞助我们！

蹲人 ！！！ 蹲算力 ！！！

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。

组织介绍

【OpenLLMAI】相信开源的力量：我们有自己的组织了！任重道远，行则将至！ - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/647882819

群组介绍：

OpenLLMAI 目前有 3 个群：

面向群友及广大的 LLM 技术爱好者：

- OpenLLM 技术交流群：无门槛，只要对 LLM/NLP 等技术有兴趣就可以申请加入（恶意引流、打广告者除外）。其中，QQ 群（无精力运营）主要负责引导大家入群，入群后请私聊管理员加入微信群。

面向正式的组织成员：

我们鼓励开源协作，所以对于正式的组织成员会有一定的门槛，除了初创成员和目前已有的成员以外，暂时只接纳对 OpenLLMAI 做出过实际贡献的同学。

- OpenLLMAI 开发者群：为了保证开发效率和质量，实行申请/邀请制，对开发工作做出实际贡献者可以私聊群主或管理员申请加入，现有成员也可以邀请相关的开发者加入。
- OpenLLMAI 研究者群：为了保证更高质量的技术交流和研究需求（组织后面也会有这方面的产出），实行申请/邀请制，对 OpenLLMAI 做出实际贡献者可以私聊群主或管理员申请加入，现有成员也可以邀请相关的开发者加入。

贡献方式：

开发：

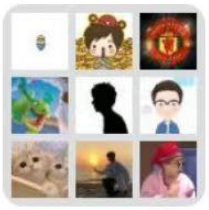
- 直接在 GitHub 上认领相关任务，如果是全新的需求，可以先提 issue，然后找 reviewers 确认是否有必要做。完成 1 次有效的 PR 后（需要有一定的代码量，不能纯为 PR 而 PR，比如修改了一个 print 语句之类的）可以申请加入 OpenLLMAI 开发者群。

其他贡献方式：以下任何一种方式，均可加入 OpenLLMAI 研究者群

- 组织一次面向群友的技术分享：技术专题、论文等等
- 主持和编辑一次 OpenLLM Talk
- 组织一次头脑风暴
- 科研协作：有科研想法想找人合作的可以找群主/管理员私聊，确认之后可以加入研究者群。

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