OpenLLM Talk 003

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 003 (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】：20230708 晚上九点（每周六晚上九点，节假日顺延）

【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；vllm 讨论；

【本期贡献者】- 排名不分先后：

【主持人】：suc16、初七（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】：【【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

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】：羡鱼、suc16

【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

LongNet: Scaling Transformers to 1,000,000,000 Tokens

- 动机：在大型语言模型时代，扩展序列长度已经成为一个关键需求。然而，现有方法在计算复杂性和模型表达性之间的平衡上存在挑战，限制了最大序列长度。本文的目标是介绍一种可以将序列长度扩展到超过 10 亿 Token 的 Transformer 变体，而不会牺牲较短序列的性能。
- 方法：提出 LONGNET，一种使用新的组件——扩张注意力(dilated attention)替换标准 Transformer 注意力的方法。扩张注意力的设计原则是随着 Token 之间距离的增长，注意力分配呈指数级下降。LONGNET 具有线性的计算复杂性和对 Token 之间的对数依赖性，可以解决有限的注意力资源和每 Token 可访问性之间的矛盾。
- 优势：该方法能有效地处理长序列，并且在各种任务中都显示出了其有效性。此外，该方法可以并行训练，打破了计算和内存的限制，使得序列长度可以有效地扩展到 10 亿 Token。

from-https://hub.baai.ac.cn/view/27697

项目

北大 chatlaw

LLM 打辩论 https://github.com/Skytliang/Multi-Agents-Debate

https://github.com/InternLM/InternLM

上海 ai lab 的 7b llm 开源了

工业界

首测生成、多轮对话能力！SuperCLUE-Open 中文大模型开放域测评基准发布

https://mp.weixin.qq.com/s/hSWfkkWmQ0rmVPTucJh5zg

文心大模型升级 3.5 版本，有多强？我们帮你试了试

盘古大模型

https://mp.weixin.qq.com/s/MYcnyG9vcw831hfkl58xlw

华为云盘古大模型登 Nature：秒级完成气象预测，速度快 10000 多倍

https://mp.weixin.qq.com/s/MYcnyG9vcw831hfkl58xlw

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】：

资料

《大语言模型综述》

https://arxiv.org/pdf/2303.18223.pdf

论文链接：https://arxiv.org/abs/2303.18223

GitHub 项目链接：https://github.com/RUCAIBox/LLMSurvey

中文翻译版本链接：

https://github.com/RUCAIBox/LLMSurvey/blob/main/assets/LLM_Survey__Chinese_V1.pdf

苏神的新博客，解读 rope

https://spaces.ac.cn/archives/9675

NTK-Aware Scaled RoPE 原文：

https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/

longchathttps://huggingface.co/lmsys/longchat-13b-16k

https://sota.jiqizhixin.com/home

RLHF 文本生成图模型 - 远洋之帆的文章 - 知乎

https://zhuanlan.zhihu.com/p/641652465

模型

百川模型信息抽取时**自动纠错**，巨硬 微软

https://mp.weixin.qq.com/s/kxnIB62uZk52IGQnjIP3Ew

浦江书生？推理速度超快

https://github.com/InternLM/InternLM

2.9M RWKV 超小模型可以搞简单数学计算。https://github.com/BlinkDL/RWKV-LM/tree/main/RWKV-v4neo/math_demo

~全网最小（？）的语言模型，适用于个人开发者在普通电脑上的训练和微调。

项目

杂项

示例：可参考 OpenLLM Talk 001 中的 state of gpt

GLM 讲座

https://m.bilibili.com/video/BV1zM4y1j7zt?buvid=Z048450D293B33244152B5D5EE88703A440B&is_story_h5=false&mid=89dsuzXgkP558Jr%2FdLF4Iw%3D%3D&p=1&plat_id=114&share_from=ugc&share_medium=iphone&share_plat=ios&share_session_id=CA185D3A-4040-4495-AAD3-D00AE55E9314&share_source=WEIXIN&share_tag=s_i&timestamp=1688803945&unique_k=2Z19cPm&up_id=503316308

何枝分享

https://m.bilibili.com/video/BV1a14y1o7fr?buvid=Z048450D293B33244152B5D5EE88703A440B&is_story_h5=false&mid=89dsuzXgkP558Jr%2FdLF4Iw%3D%3D&p=1&plat_id=114&share_from=ugc&share_medium=iphone&share_plat=ios&share_session_id=8705C858-6F45-4063-B8BA-876205A0C797&share_source=WEIXIN&share_tag=s_i&timestamp=1688646095&unique_k=OrUWwGk&up_id=507524288

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】：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】：

【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**；

```
old_init =
transformers.models.llama.modeling_llama.LlamaRotaryEmbedding.__i
nit__
def adaptive_ntk_init(self, dim, max_position_embeddings=2048,
base=10000, device=None):
    self.dim = dim
    self.base = base
    old_init(self, dim, max_position_embeddings, base, device)

def adaptive_ntk_forward(self, x, seq_len=None):
    if seq_len > self.max_seq_len_cached:
        t = torch.arange(seq_len, device=x.device,
dtype=self.inv_freq.dtype)
        inv_freq = self.inv_freq
        dim = self.dim
        alpha = seq_len / 1024 - 1
        base = self.base * alpha ** (dim / (dim-2))
        inv_freq = 1.0 / (base ** (torch.arange(0, dim,
2).float().to(x.device) / dim ))
        freqs = torch.einsum("i,j->ij", t, inv_freq)
        emb = torch.cat((freqs, freqs), dim=-1).to(x.device)
        cos_cached = emb.cos()[None, None, :, :]
        sin_cached = emb.sin()[None, None, :, :]
        return (
            cos_cached[:, :, :seq_len, ...].to(dtype=x.dtype),
            sin_cached[:, :, :seq_len, ...].to(dtype=x.dtype)
        )
    return (
        self.cos_cached[:, :, :seq_len, ...].to(dtype=x.dtype),
        self.sin_cached[:, :, :seq_len, ...].to(dtype=x.dtype)
    )
transformers.models.llama.modeling_llama.LlamaRotaryEmbedding.for
ward = adaptive_ntk_forward
transformers.models.llama.modeling_llama.LlamaRotaryEmbedding.__i
nit__ = adaptive_ntk_init
```

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 只有几百个中文 token，那么对于一些不在这些 token 中的汉字是如何通过多个 token 表示的？参考信息: sentence piece 使用的 unicode 编码?
    a. LLaMA 的 sp model 应该是更接近于 BBPE 的方式，byte-level 编码
2. ChatGPT 的数学能力的来源是什么呢，比如一个非常大的数字乘以一个非常大的数字，ChatGPT 虽然给不出一个比较准确的答案，但是能回答出一个差不多的结果。（A: 这和 ChatGPT 没有关系。本质上是一个计算复杂度问题，比如你考虑更一般的多项式乘法，f(x) = a_0x^0+...+a_nx^n, g(x) = b_0x^0+...+b_mx^m，那么前几项 a_n b_m，(a_{n-1}b_m + a_n b_{m-1}), ...和后几项 a_0b_0, a_1b_0+a_0b_1, a_2b_0+a_1b_1+a_0b_2, ...都是很容易算出来的，但是中间几项不好算。十进制乘法不过是 x=10 的特殊情况。模型由于一次输出一个词元，本质上是以 O(n)的复杂度计算 O(n logn)甚至 O(n^2)的乘法，因此不可能算准的。）

群里讨论：

有空会同步，取决于人力，希望大家积极认领~

1. vllm:

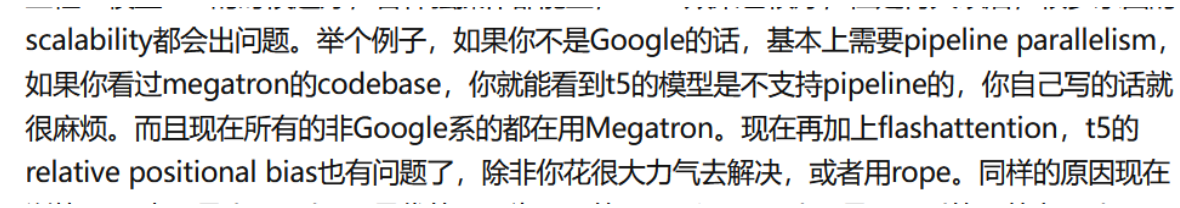
vllm 有一个 gpu_memory_utilization(显存利用率)参数，先加载模型到内存中，然后获取当前显卡最大内存容量*显存利用率-当前模型内存，得到的 size 就是用来做 cache 的内存。输入的 context 增加，消耗的内存基本上也会从 cache 中取，不会消耗更多。

还有个默认的 swap 参数，默认为 4，会在 CPU 中分配 4 个 G 的 CPU 内存做 cache，每个 block_cache_size 大概是 80M 还是多少来着，然后会根据上面的 GPU 和 CPU 内存缓存量/cache_size，得到 GPU 和 CPU 的 cache_block 数目。CPU 默认好像是 512 个，GPU 的 block 数目根据 gpu_memory_utilization 参数会调整.

vllm 很快的，主要是调度优化和节省了 kv cache 显存，batchsize 可以开得更大，所以

快

2. 为什么 LLM 都是基于 decoder-only，这个能讨论下？



【OpenLLM 001】大模型的基石-架构之争，decoder is all you need? - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/621192550

3. 有没有用 lora 微调、p-tuning v2 微调，基于自身项目上实验效果比 bert 全量微调好的？
4. 模型的 delta 权重是啥意思呀，看一些基于 llama 的 demo 放出来的是 delta 权重，是不是说明这模型是基于全量微调获得的

参考资料

《大语言模型综述》

https://arxiv.org/pdf/2303.18223.pdf

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，字符检索+语义检索，P0；

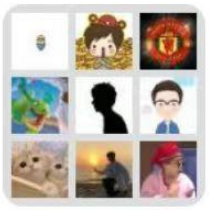
- OpenLLM：LLM 学习和实践项目，P0；
- OpenAIDic：科普项目；
- ChatLover：模拟恋人+爱情助手，P1；

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 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