

大模型实战营

书生浦语作业

第一节笔记

上海人工智能实验室目前开放了大模型的全链条开源体系，涵盖数据集、预训练、微调、部署、评测、应用等模块，对从事大模型相关工作、研究的同学比较友好，消费级的显卡也可以训练大模型



书生包含大量的数据

融合Hybird Zero技术，可大大加速训练

全链条开源开放体系 | 预训练

高可扩展	训练算法	预训练	微调
支持从 8 卡到千卡训练，千卡加速效率达 92%	训练优势	高性能 Transformer 计算库	多种并行策略
极致性能优化	通信/计算调度	梯度累积算法选择	通信/计算重叠
兼容主流	显存管理	优化器状态	梯度
开箱即用			参数

卡均训练吞吐量 @512 卡 (tokens/gpu/s)

框架	吞吐量 (tokens/gpu/s)
Megatron-deepspeed	1774
通义训练框架	3835

训练加速效率

OpenMMLab 2024/01/01 22:38:54

开发XTuner微调框架，适配HuggingFace等主流开源模型库，并自带加速库，适配多种硬件

全链条开源开放体系 | 微调

高效微调框架 XTuner

适配多种生态

- 多种微调算法
- 适配多种开源生态
- 自动优化加速

适配多种硬件

- 训练方案覆盖 NVIDIA 20 系以上所有显卡
- 最低只需 8GB 显存即可微调 7B 模型

一个规范标准的大模型评测体系可以更好的帮助我们对大模型的能力进行全面的测评

全链条开源开放体系 | 评测

国内外评测体系的整体态势

机构	Stanford University	北京智源人工智能研究院	Berkeley	Stanford University	CLUE	Hugging Face
类型	客观评测	客观/主观评测	客观评测	主观评测	客观/主观评测	客观评测
量级	5W+ 英文题目	8W+ 中英双语	1W+ 英文题目	1K+ 英文题目	3K+ 中文题目	2W+ 英文题目

 学科	 语言	 知识	 理解	 推理	 安全
初中考试	字词释义	知识问答	阅读理解	因果推理	偏见 有害性
中国高考	成语习语	多语种知识问答	内容分析	常识推理	公平性 隐私性
大学考试	语义相似		内容总结	代码推理	真实性 合法性
语言能力考试	指代消解			数学推理	
职业资格考试	翻译				

模型部署也融合了一些主流的推理优化技术，本人对这块比较感兴趣，希望可以深入研究一下

大语言模型特点	技术挑战	部署方案
内存开销巨大 <ul style="list-style-type: none"> 庞大的参数量 采用自回归生成token，需要缓存k/v 动态Shape <ul style="list-style-type: none"> 请求数不固定 token逐个生成，且数量不定 模型结构相对简单 <ul style="list-style-type: none"> transformer 结构，大部分是 decoder-only 	设备 <ul style="list-style-type: none"> 低存储设备（消费级显卡、移动端等）如何部署？ 推理 <ul style="list-style-type: none"> 如何加速 token 的生成速度 如何解决动态shape，让推理可以不间断 如何有效管理和利用内存 服务 <ul style="list-style-type: none"> 提升系统整体吞吐量 降低请求的平均响应时间 	技术点 <ul style="list-style-type: none"> 模型并行 低比特量化 Attention优化 计算和访存优化 Continuous Batching