

1. 大模型成为发展AGI的重要途径



2. InternLM开源了7B大模型、语料以及全链路开发工具



4. 数据平台



5. 模型微调以及提供了微调工具Xtuner

全链条开源开放体系 | 微调

大语言模型的下游应用中，增量续训和有监督微调是经常会用到两种方式。

增量续训

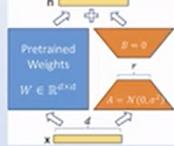
使用场景：让基座模型学习到一些新知识，如某个垂类领域知识
训练数据：文章、书籍、代码等

部分参数微调

有监督微调

使用场景：让模型学会理解和遵循各种指令，或者注入少量领域知识
训练数据：高质量的对话、问答数据

全量参数微调



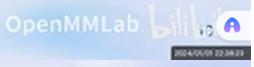
高效微调框架 Xtuner

任务类型	数据格式	训练引擎	优化加速	支持算法
增量预训练 指令微调 工具类指令微调	Alpaca MOSS OpenAI Guanaco	MV Engine	Flash Attention DeepSpeed ZeRO Pytorch FSDP	QLoRA 微调 LoRA 微调 全量参数微调
		消费级显卡 GeForce RTX 2080、2080Ti GeForce RTX 3060 ~ 3090Ti GeForce RTX 4060 ~ 4090	数据中心 Tesla T4、V100 A10、A100、H100	适配多种生态

- 多种微调算法
多种微调策略与算法，覆盖各类 SFT 场景
- 适配多种开源生态
支持加载 HuggingFace、ModelScope 模型或数据集
- 自动优化加速
开发者无需关注复杂的显存优化与计算加速细节

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



6. 开源了评测工具OPENCOMPASS

全链条开源开放体系 | 评测

OpenCompass 开源评测平台架构



工具层

分布式评测 提示词工程 评测数据库上报 评测榜单发布 评测报告生成

方法层

自动化客观评测 基于模型辅助的主观评测 基于人类反馈的主观评测

能力层

通用能力：学科、语言、知识、理解、推理、安全
特色能力：长文本、代码、工具、知识增强

模型层

基座模型 对话模型



7. 大模型部署中的挑战和LMDeploy开源工具



全链条开源开放体系 | 部署

大语言模型特点

技术挑战

部署方案

内存开销巨大

- 庞大的参数量
- 采用自回归生成token，需要缓存k/v

动态Shape

- 请求数不固定
- token逐个生成，且数量不定

模型结构相对简单

- transformer结构，大部分是decoder-only

设备

- 低存储设备（消费级显卡、移动端等）如何部署？

推理

- 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

- 提升系统整体吞吐量
- 降低请求的平均响应时间

技术点

- 模型并行
- 低比特量化
- Attention优化
- 计算和访存优化
- Continuous Batching

LMDeploy 提供大模型在GPU上部署的全流程解决方案，包括模型轻量化、推理和服务。

8. 多模态智能体工具箱AgentLego可以让LLM和环境结合起来



全链条开源开放体系 | 智能体

多模态智能体工具箱 AgentLego

- 丰富的工具集合，尤其是提供了大量视觉、多模态相关领域的前沿算法功能
- 支持多个主流智能体系统，如 LangChain, Transformers Agent, Lagent 等
- 灵活的多模态工具调用接口，可以轻松支持各类输入输出格式的工具函数
- 一键式远程工具部署，轻松使用和调试大模型智能体

Hugging Face
OpenMMLab
Stable Diffusion
SAM

LangChain
Lagent
Transformers Agents

OpenMMLab 2024/05/01 22:55:07