11 大模型部署

1 大模型特点

内存开销巨大
庞大的参数量。7B模型仅权重就需要14+G内存
采用自回归生成token,需要缓存Attention的kN,带来巨大的内存开销

动态shape
请求数不固定
Token逐个生成，且数量不定

相对视觉模型，LLM结构简单
Transformers结构，大部分是decoder--only

---

2 模型部署

定义
a.将训练好的模型在特定软硬件环境中启动的过程，使模型能够接收输入并返预测结果
b.为了满足性能和效率的需求，常常需要对模型进行优化，例如模型压缩和硬件加速

产品形态
云端、边缘计算端、移动端

计算设备
CPU、GPU、NPU、TPU等

---

3 大模型部署挑战

设备
如何应对巨大的存储问题？低存储设备（消费级显卡、手机等)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
如何提升系统整体吞吐量?
对于个体用户，如何降低响应时间?

---

4 大模型部署方案

技术点
a.模型并行
b.transformer计算和访存优化
c.低比特量化
d.Continuous Batch
e.Page Attention

方案
a.huggingface transformers
b.专门的推理加速框架