从大模型推理极限理论最优值谈谈推理优化

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年08月11日 10:20 上海

◈ 技术总结专栏 ◈

作者：喜欢卷卷的瓦力

本篇基于大模型推理机制及其极限理论值，讲述其具体在推理优化中起到的作用

之前的文章 大模型推理瓶颈及极限理论值分析 分析了大模型推理的速度瓶颈及量化评估方式，本文来谈谈用途，希望对小伙伴们理解大模型推理内部工作机制与推理优化有帮助。

根据上篇内容可以很容易地计算出推理所需的最小时间，下面是从参考资料中找到的一些单卡推理测试示例（16bit），下面是具体的平均延迟，仅供参考：

LLM	RTX 4090 24GB (2022)	A100 80GB (2020)	V100 32GB (2017)
ChatGLM3-6B	16ms/token	18ms/token	32ms/token
Qwen-7B	19ms/token	29ms/token	41ms/token

接下来看看这些理论极限有什么用。

下面是一个快捷目录。

用途1：评估推理系统好坏
用途2：指导量化
用途3：指导模型优化方向
用途4：硬件相对推理速度评估

用途1：评估推理系统好坏

要接近理论极限，需要一个高质量的软件实现，以及能够达到峰值带宽的硬件。

因此如果你的软件+硬件最终得到的结果离推理的理论最优很远，那肯定就有问题；可能需要具体排查硬件问题还是软件问题。

例如，在 RTX 4090 上 calm 使用 16 位权重时达到 ~15.4 ms/tok，使用 8 位权重时达到 ~7.8 ms/tok， 达到了理论极限的 90%。

用途2：指导量化

带宽与每个权重使用的 bit 数成正比；这意味着**更小的权重格式（量化）上延迟更低**。例如，在 RTX 4090 上 llama.cpp 使用 Mistral 7B

- 16 bit 权重：~17.1 ms/tok（82% 的峰值）
- 8.5 bit 权重：~10.3ms/tok （71% 的峰值）
- 4.5 bit 权重：~6.7ms/tok （58% 的峰值）

因此对于低延迟场景，可以考虑低精度量化。

用途3：指导模型优化方向

推理估算还表明了 推理过程并未充分利用算力（ALU）。要解决这个问题，需要重新平衡 FLOP:byte 比例， [speculative decoding] 等技术试图部分解决这个问题。

1）推理时的batch size扩展：从1—>N

这种场景下，瓶颈不再是访存 IO，而是算力（ALU）：

- 当多个用户请求同时处理时，用相同的矩阵同时执行多个矩阵-向量乘法， 可以将多个矩阵-向量乘法变成一个矩阵-矩阵乘法。
- 对于足够大的矩阵来说，只要矩阵-矩阵乘法实现得当，速度就比访存 IO 快，

因此这种场景下，瓶颈不再是访存 IO，而是算力（ALU）。
这就是为什么这种 ALU:byte 不平衡**对于生产推理系统不是关键问题**—— 当使用 ChatGPT 时，你的请求与同一 GPU 上许多其他用户的请求并发评估，GPU 显存带宽利用更加高效。

2) 批处理无法改善所需加载的 KV-cache 数据量

批处理通常不会减轻 KV-cache 带宽（除非多个请求共享非常大的前缀），因为 KV-cache 大小和带宽**随请求数量的增加而增加**，而不像权重矩阵保持不变。

用途4：硬件相对推理速度评估

带宽是评估推理性能的关键指标，对于模型变化/设备类型或架构来说是一个恒定的， 因此即使无法使用 batch processing，也可以用它来评估你用的硬件。

例如，NVIDIA RTX 4080 有 716 GB/s 带宽，所以可以预料到它的推理速度是 RTX 4090 的 ~70% —— 需要注意的是，游戏、光线追踪或推理其他类型的神经网络等方面，相对性能可能与此不同。

想要获取技术资料的同学欢迎关注公众号，进群一起交流~

瓦力算法学研所
我们是一个致力于分享人工智能、机器学习和数据科学方面理论与应用知识的公众号。我...
117篇原创内容

公众号

参考文献

[1] LLM inference speed of light—（https://zeux.io/2024/03/15/llm-inference-sol/）
[2] ArthurChiao's blog —（https://arthurchiao.art/blog/llm-inference-speed-zh/）

学术理论解析 53

学术理论解析 · 目录

上一篇
大语言模型在生成式信息提取中的应用概览

下一篇
Megatron-LM，又一大模型训练神器