AdaptiveLLM: 基于张量交换和张量重算的大语言模型推理优化技术

段晓辉[1]
[1] 清华大学，北京 100084
(sunrise_duan@126.com)

AdaptiveLLM: Efficient LLM inference based on swapping and re-computation

Xiaohui Duan[1]
[1] (Tsinghua University, Beijing 100084)

Abstract Large Language Models(LLMs) come with an extremely high amount of parameters, posing significant challenges for inference tasks. Traditional LLM inference services employs swapping and re-computation techniques, guaranteeing the success of the inference tasks at the cost of performance on limited GPU memory. However, existing LLM serving systems fails to search memory management schemes adaptively based on the runtime information of LLM inference tasks, leading to a sub-optimal performance. And at the same time, these works are inferior in holding a balance between inference throughput and request latency, targeting at only one in these two objectives in the experiments and neglecting the other. To address the above issues, we propose AdaptiveLLM, an efficient LLM service for inference tasks based on swapping and re-computation technique. Fundamentally, we implement an overhead predictor for swapping and re-computation, with an error rate lower than 2% and 4% respectively. Moreover, we develop a cost-aware memory optimization method and a fairness-based request scheduling algorithm. The former speeds the inference task on the server, and the latter improves the real-time performance by reducing request latency targeting at the client. We conduct experiments on typical LLMs and datasets while setting vLLM as the baseline. As a result, the cost-aware memory optimization method accelerates the inference task by 1.1 to 1.4, and the fairness-based request scheduling algorithm can significantly reduce the average weighted around time by 60% to 80%. This demonstrates that AdaptiveLLM makes a better trade off between throughput and latency by resolving their discrepancy in implementation, and achieves efficient LLM inference.
Key words LLM, inference, swapping, re-computation

摘要 大语言模型（*LLMs*）拥有极高的参数量，为推理任务的高效运行带来巨大挑战。传统的 *LLM* 推理框架引入了张量交换和张量重算等技术，在有限的 *GPU* 内存上以牺牲性能为代价完成 *LLM* 推理。然而，已有研究工作无法根据推理任务运行时信息自适应地选择内存优化技术，导致推理任务的性能无法得到进一步提升；同时这些工作没有实现整体吞吐率与单请求延时之间的权衡，仅能以二者之一作为优化目标。针对以上问题，本文面向大模型推理服务场景，提出了 *AdaptiveLLM*，一款基于张量交换和张量重算的 *LLM* 推理框架。首先，*AdaptiveLLM* 实现了张量重算和张量交换开销预测，其预测误差分别在 *2%* 和 *4%* 以下。其次，该框架引入了基于开销感知的张量优化策略，旨在实现面向服务器端的处理加速；同时引入了基于公平性的用户请求调度策略，旨在实现面向客户端的实时请求处理。本文在常见 *LLM* 和推理数据集上开展实验，并将 *vLLM* 框架作为基础程序。结果表明，基于开销感知的张量优化策略能够为推理任务带来 *1.1* 到 *1.4* 的整体吞吐加速比；基于公平性的用户请求调度策略能够降低平均带权周转时间为 *60%* 至 *80%*。由此证明 *AdaptiveLLM* 在优化过程中权衡整体吞吐率与单请求延时，化解了二者在实现上的冲突，实现 *LLM* 高效推理。

关键词 大语言模型、推理、张量交换、张量重算

中图法分类号 TP391

基金项目：目前无人捐助此项目