AdaptiveLLM：基于自适应张量交换和张量重算的大语言模型推理优化

梁绪宁[1]，王思琪[1]，杨海龙[1]，栾钟治[1]，刘轶[1]，钱德沛[1]

[1] 北京航空航天大学，北京 100191

(liangxuning@126.com,lethean1@buaa.edu.cn,hailong.yang@buaa.edu.cn,07680@buaa.edu.cn,yi.liu@buaa.edu.cn,depeiq@buaa.edu.cn)

AdaptiveLLM: Efficient LLM inference using adaptive tensor swapping and re-computation techniques

Liang Xuning[1]，Wang Siqi[1]，Yang Hailong[1]，Luan Zhongzhi[1]，Liu Yi[1]，Qian Depei[1]

[1] (Beihang University, Beijing 100191)

Abstract Large Language Models(LLMs) come with an extremely high amount of parameters, posing significant challenges for inference tasks. Traditional LLM inference services employ tensor swapping and tensor re-computation techniques, guaranteeing the success of generation at the cost of performance on limited GPU memory. However, existing LLM serving systems fail to search memory management schemes adaptively based on runtime information, leading to a sub-optimal performance. Furthermore, these works are inferior in the trade-off between throughput and latency, preferring on only one and compromising the other. To address the above issues, we propose *AdaptiveLLM*, an efficient LLM serving framework for inference tasks based on swapping and re-computation. Specifically, *AdaptiveLLM* implements an overhead predictor for swapping and re-computation, with an error rate lower than 2% and 4% respectively. *AdaptiveLLM* also adopts a cost-aware memory optimization algorithm which improves throughput on the server, and a fairness-based request scheduling algorithm which reduces latency for the client. On typical LLMs and datasets, our evaluation shows that *AdaptiveLLM* improves the throughput by 10% to 40%, and reduces the average weighted around time by 20% to 40%, compared with the vLLM baseline. In conclusion, *AdaptiveLLM* achieves efficient LLM inference by making a better trade off between throughput and latency.

Key words LLM; Inference; Tensor Swapping; Tensor Re-computation; Adaptive Memory Optimization

摘要 大语言模型（*LLMs*）拥有极高的参数量，为推理任务带来 *GPU* 内存瓶颈。已有 *LLM* 推理框架引入张量交换和张量重算等内存优化技术，在有限的 *GPU* 内存上通过牺牲性能完成推理。然而，已有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进一步提升。同时，这些工作以推理任务整体吞吐率或单请求响应延时为单一优化目标，缺乏对上述优化目标的综合考虑。针对以上问题，本文面向 *LLM* 推理服务场景，提出了 *AdaptiveLLM*，一款基于自适应张量交换和张量重算的 *LLM* 推理框架。*AdaptiveLLM* 实现了张量重算和张量交换开销精准预测，其预测误差分别控制在 *2%* 和 *4%* 以下。在此基础上，引入了基于开销感知的内存优化策略，可以自适应地选择开销较低的内存优化技术，提高了任务整体吞吐率。同时，引入了基于公平性的用户请求调度策略，降低了单请求延时。本文在主流 *LLM* 模型和数据集上开展实验验证，以 *vLLM* 作为基准程序进行对比评估。结果表明，*AdaptiveLLM* 实现了 *10%* 到 *40%* 的整体吞吐率提升，同时使平均带权周转时间降低了 *20%* 至 *40%*。由此证明 *AdaptiveLLM* 在推理优化过程中可以更好地权衡整体吞吐率与单请求延时，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；自适应内存优化

中图法分类号 TP391

基金项目：目前无人捐助此项目

1. 引言

从人脸识别 [1]、个性化推荐 [2]、到智能家居 [3]、无人驾驶 [4] 等应用领域，深度学习（Deep Learning, DL） [5] 相关技术已经融入到社会的方方面面，为人类的生产生活带来了极大的便利。自然语言处理（Natural Language Processing, NLP） [6] 作为深度学习领域的重要研究方向，长期以来备受研究人员关注。近年来，随着 GPU 算力的不断提升，各种语言模型也朝着更大参数量、更高准确度的方向迅猛发展。

大语言模型 [7]（Large Language Models，LLM）是自然语言处理领域的一个分支。LLM 通常拥有十亿级别，甚至万亿级别的参数量，因此需要海量文本数据进行训练。同时，LLM 在多种类型的任务中展现出卓越性能，如文本摘要 [8]、机器翻译 [9]、代码生成 [10]、以及对话问答 [11] 等，拥有巨大的科研潜力与商业价值。2021 年 GPT-3 模型 [8, 12] 的问世标志着 LLM 研究领域的一个里程碑，自此，各大科研机构纷纷投入到相关研究中，各种 LLM 层出不穷，使得该领域的研究和应用热度空前高涨。

复杂的模型结构和庞大的参数量为 LLM 带来了卓越的应用效果，但却为 LLM 部署后的推理性能优化带来了极大挑战。特别地，LLM 庞大的参数量导致推理过程中产生极高的内存占用。例如，GPT-175B 模型仅在权重加载环节就需要消耗 325GB 的 GPU 内存空间 [13]，需要使用至少 5 个 NVIDIA A100 GPU（80GB），并引入复杂的推理并行化策略才能够完成模型推理。因此，如何降低 LLM 推理任务的内存资源占用对于 LLM 成功部署和推广至关重要。

传统的深度学习模型研究中提出了张量交换 [14]（Swapping）、张量重算 [15]（Recomputation）等技术来降低推理过程中的显存占用。具体而言，张量交换技术通过张量生命周期分析、异步传输、动态调度等机制，将推理过程中不需要立即使用的张量从 GPU 内存交换到 CPU 内存；而张量重算则是将推理过程中的部分中间张量在不需要时释放，将计算图的关键节点保存下来，并在需要时重新计算这些中间结果。然而，简单地将张量交换和张量重算技术应用于 LLM 推理框架会导致以下两点不足：

首先，张量交换和张量重算技术虽然可以降低推理过程的 GPU 显存资源占用，但其对 LLM 推理性能的影响十分复杂，取决于服务器硬件配置（如 GPU 计算能力，CPU-GPU 传输带宽）、用户设置（如生成新 token 时的采样方式）、LLM 任务类型与数据集选取、以及推理任务的运行时信息等。已有的 LLM 推理框架 [14, 16-17] 虽然已经集成了张量交换或张量重算技术，但其在 GPU 显存不足时只能固定选择上述技术中的一种，而无法根据上述信息选择更优者，显著影响推理任务的性能。

另外，单请求延时和整体吞吐率在优化过程中存在矛盾。一部分传统工作，如 FasterTransformer [18] 等，以单请求延时为单一优化目标，将大部分集群资源分配给当前运行的少数请求，使得集群资源利用率降低，限制整体吞吐率；另一部分工作，如 ORCA [16-17] 等，以整体吞吐率为单一优化目标，通过提升批处理大小来增加资源利用率，但会导致单请求时延增加。整体吞吐率是面向服务器端的性能优化指标，体现了服务器端的处理效率。单请求平均延时是面向客户端的性能优化指标，体现了用户请求处理的实时性。在使用张量交换或张量重算技术优化 LLM 推理任务显存占用时，需要考虑其对推理任务整体吞吐率和单请求延时的影响，并在二者间进行权衡。

为了解决上述缺陷，本文提出了 AdaptiveLLM，一个基于张量交换和张量重算的自适应 LLM 推理服务框架。该框架实现了针对张量交换和张量重算的精准开销分析，并调用基于开销感知的内存优化策略和基于公平性的用户请求调度策略，在张量交换和张量重算技术间进行动态选择，在降低 LLM 推理显存占用的同时，降低其对 LLM 推理吞吐量和单请求时延的影响，进而实现 LLM 任务的高效推理。

- **本文设计了一款张量重算开销分析器，实现了张量重算开销的精准预测。**通过算子粒度计算复杂度分析来识别张量重算开销的影响因素，而后建立回归预测模型预测单步推理执行时间。实验表明，张量重算开销分析器的预测误差在 2% 以内。
- **本文设计了一款张量交换开销分析器，实现了张量交换开销的精准预测。**本文获取用户请求 KV Cache 的内存占用和 GPU-CPU 间通信效率信息，来对张量交换的数据传输开销进行预测。实验表明，张量交换开销分析器的预测误差在 4% 以内。
- **本文设计和实现了一个基于张量交换和张量重算的自适应 LLM 推理服务框架 AdaptiveLLM。**该框架引入基于开销感知的内存优化策略，动态选择相应的内存优化技术，通过降低显存占用，来提升推理任务的整体吞吐率。同时引入基于公平性的用户请求调度策略，降低单请求平均延时和带权周转时间。
- **本文选择典型 LLM 模型和数据集对 AdaptiveLLM 进行全面实验评估。**在典型 LLM 模型（OPT [19]、Llama [20]）和数据集（Chatbot [21]、Alpaca [22]、Summary [23]）上对 AdaptiveLLM 的张量重算开销分析器、张量交换开销分析器、

内存优化决策器、和用户请求调度器模块的有效性进行了实验验证。结果表明，AdaptiveLLM 能够实现 10% 到 40% 的整体吞吐率提升，且将用户请求平均带权周转时间降低 20% 至 40%。

2. 背景知识

本章介绍有关 AdaptiveLLM 的背景知识。由于 AdaptiveLLM 主要面向 LLM 推理过程中产生的 KV Cache 内存占用进行优化，因此本章将在第一节阐明 KV Cache 在 LLM 推理任务中的功能，在第二节论述传统工作中面向 KV Cache 的内存优化技术。

2.1. KV Cache 的提出

LLM 推理任务以 token 作为输入与输出的基本单位。对于生成式推理任务，每次前向传播计算仅生成一个新 token。一般来说，其包含两个阶段：prefill 阶段读取用户输入的 token 序列，生成第一个 token；decode 阶段分为多步进行，依次生成后续 token，直至得到终止 token。在推理过程中，每个 token 拥有一个 key-value 张量对，为自注意力机制下的编码结果。

在 decode 阶段中，每个 token 的计算均依赖于前序 token 的 key 值和 value 值。如果每次计算前都重新调用自注意力机制来获取前序 token 的 key-value 张量，则会产生大量不必要的计算开销。主流 LLM 推理服务 [14, 16-17, 24-25] 框架普遍采用 KV Cache 数据结构来保存这些 token 的 key-value 张量，方便后续 token 的生成，避免重复计算。

然而，随着后续 token 的不断生成，KV Cache 迅速扩展，产生推理内存瓶颈。例如，在 OPT-13B 模型中，对于一个长度为 100 的用户请求，其 KV Cache 能够占用 39.1MB 的内存空间。有限的 GPU 内存将批处理大小限制在较低水平，阻碍推理并发度的进一步提升，进而限制吞吐率。

不同于 LLM 参数张量，KV Cache 占用的内存空间在对应用户请求推理完毕后被释放。其内存占用量大、动态性高，拥有较大的优化空间，因此 AdaptiveLLM 的内存优化策略将针对 KV Cache 实现。

2.2. KV Cache 的内存优化

KV Cache 的引入方便了计算过程，却带来内存瓶颈，使得 LLM 推理性能的提升无法达到预期水平。下面介绍针对 KV Cache 内存占用的一些优化工作。

(1) 内存碎片优化

在传统 LLM 推理服务框架 [14] 中，内存管理器按照用户定义的序列长度上限，为每个请求设置一块固定大小的 GPU 内存来存储 KV Cache。但用户请求长度的差异性导致内碎片的大量产生。为了解决该问题，部分 LLM 推理框架 [26] 能够基于历史信息来预测输出长度，并按照预测值分配内存。然而，预测误差会导致输出截断，且旧请求的完成与新请求的加入使得内存中产生很多外碎片。随着新请求的不断到来，内碎片与外碎片在内存中积累，严重影响了内存空间的高效使用。基于这些问题，vLLM 框架 [16] 引入了 Paged Attention 机制，基于 OS 页式内存管理思想，将 GPU 内存划分成块，并通过维护块表来支持 KV Cache 在内存空间中的不连续存储。该机制基本消除了内碎片和外碎片现象，大大提升内存利用率。

(2) 张量交换与张量重算

为了攻克推理内存瓶颈，传统框架引入了张量交换技术 [14, 16, 27]，将暂时不会使用的 KV Cache 传输至 CPU 中，在计算需要时重新传输至 GPU 中。然而，CPU-GPU 间有限的 PCIe 带宽使得换出和换入过程产生不可忽略的通信开销，限制吞吐率，降低推理性能。部分研究提出 [15]，当张量交换带来的开销超过重新调用自注意力机制的开销时，应选择后者来获取所有前序 token 的 key-value 张量，也称张量重算。具体来说，内存管理器直接删除重算请求对应的 KV Cache，在其被调度时执行一次 prefill 阶段来代替原本应该执行的 decode 阶段。重算与交换的联合使用缓解了通信开销问题，然而，当 GPU 内存不足时，如何在二者中进行选择成为了新的困境。AdaptiveLLM 针对此问题设计了基于开销感知的内存优化策略，能够预测二者的开销，并选择开销小的过程执行。

3. AdaptiveLLM 的设计与实现

本章介绍了本文工作的具体设计。第一节给出 AdaptiveLLM 的整体设计方案，后面的章节将分别介绍 AdaptiveLLM 中不同的功能模块。

3.1. 整体架构

AdaptiveLLM 实现了三个主要功能模块，包括张量重算开销分析器、张量交换开销分析器和自适应 LLM 推理优化器，其整体架构如图1所示。

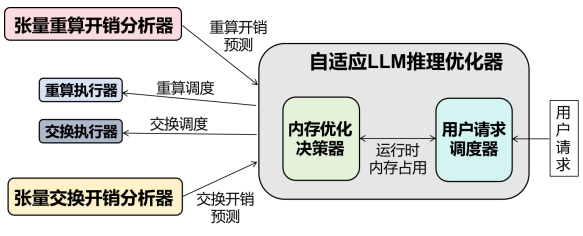

图 1 整体设计架构

张量重算开销分析器基于序列长度、模型隐藏维

度、模型层数和批处理大小来预测重算开销。张量交换开销分析器基于 KV Cache 内存占用和 GPU-CPU 双向传输带宽来预测交换开销。自适应 LLM 推理优化器包含内存优化决策器和用户请求调度器。当 GPU 内存不足时，内存优化决策器引入基于开销感知的内存优化策略，选择优先级最低的用户请求，收集张量重算开销分析器与张量交换开销分析器提供的开销预测值。选择开销小的内存优化方式，而后交付相应的执行器。该过程称为“抢占调度”。当 GPU 内存空余时，用户请求调度器使用基于公平性的用户请求调度策略，在满足公平性的前提下尽可能多地调度剩余用户请求，避免 GPU 资源浪费。该过程称为“启动调度”。推理过程中，内存优化决策器与用户请求调度器共享 KV Cache 实时内存占用信息。二者高效协同，实现整体吞吐率与单请求延时的权衡。

3.2. 张量重算开销分析器

抢占调度时，重算执行器在内存中删除用户请求的 KV Cache 张量。启动调度时，执行一次 prefill 阶段来恢复被删除的数据。因此，张量重算引入的额外开销等于被抢占请求执行 prefill 阶段的时间。

本文以 OPT 和 Llama 模型为例，通过算子粒度复杂度分析来识别单步推理时间的影响因素。

(1) 算子粒度开销分析

OPT 和 Llama 模型中包含 5 种不同的算子：ReLU、Norm、Linear、SiluAndMul 和 Attention，其计算流程如图2所示。图中 X_i,Y_i 是由用户输入决定的张量维度；$input\_dim$,$output\_dim$,$head\_size$ 是由算子本身决定的张量维度。

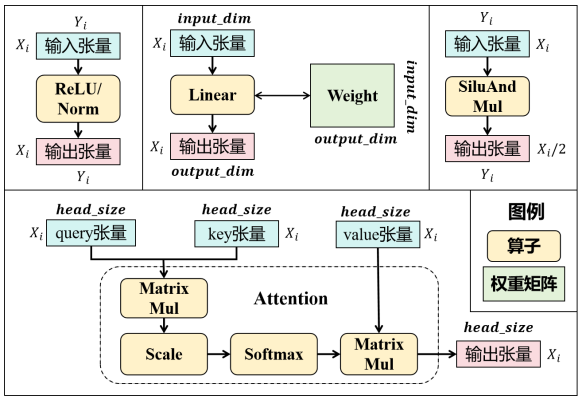

图 2 四种算子的计算流程

下面分别对这些算子进行复杂度分析。

- **ReLU 算子**：逐位调用激活函数进行计算，其时间复杂度为 $O(X_i * Y_i)$。
- **Norm 算子**：是 LayerNorm、RMSNorm（仅在 Llama 模型中）等多种归一化算子的统称，其时间复杂度为 $O(X_i * Y_i)$。
- **Linear 算子**：将输入向量从 $input\_dim$ 维空间映射到 $output\_dim$ 维空间中，其计算复杂度为 $O(X_i * input\_dim * output\_dim)$。
- **SiluAndMul 算子**：该算子仅出现在 Llama 模型的 MLP 层中，将输入向量的指定维度减半，其时间复杂度为 $O(X_i * Y_i)$。
- **Attention 算子**：属于复合操作，由矩阵乘法、缩放和 Softmax 激活等底层算子组成，整体计算过程如公式1，其时间复杂度为 $O(X_i^2 * head\_size)$。

$$Attention(Q,K,V) = softmax(\frac{Q \times K^T}{\sqrt{h}} \times V) \quad (1)$$

根据算子粒度复杂度分析，可以识别出 4 项有关 LLM 单步推理执行时间的影响因素，分别为：LLM 层数、LLM 隐藏维度、单请求需要处理的 token 数量、和批处理大小。

(2) 单步推理开销预测模型

单步迭代执行时间预测是一项拥有 4 个输入变量，1 个输出变量的回归预测任务。根据算子粒度时间复杂度分析可知，输出变量与输入变量之间存在多项式依赖关系。因此，本文共选用了 8 个回归模型，包括线性回归模型、决策树回归模型、随机森林回归模型、岭回归模型、套索回归模型、弹性回归模型、梯度提升回归模型、和 K-临近回归模型。针对每种回归模型，对不同的多项式拟合次数（1 到 10）进行测试。选择在测试集上预测误差最小的配置，并将其部署到 AdaptiveLLM 的张量重算开销分析器中。

3.3. 张量交换开销分析器

抢占调度时，交换执行器将用户请求的 KV Cache 从 GPU 传输到 CPU 中（换出阶段）。启动调度时，将其 KV Cache 传输回 GPU 中（换入阶段）。因此，张量交换引入的额外开销等于被抢占请求的换出时间与换入时间之和。

换出开销与换入开销的计算方式如公式2所示。

$$\begin{aligned} SwapOut\_Time &= \frac{KVCach\_Mem}{DtoH - bandwidth} \\ SwapIn\_Time &= \frac{KVCache\_Mem}{HtoD - bandwidth} \end{aligned} \quad (2)$$

其中 $DtoH - bandwidth$ 是数据从 GPU 传输到 CPU 的带宽，$HtoD - bandwidth$ 是数据从 CPU 传输到 GPU 的带宽。AdaptiveLLM 继承了 vLLM 所采用

的 Paged Attention 技术，在 GPU 和 CPU 内存中划分大小固定的 Block，用于存储 KV Cache。每个 Block 的内存占用如公式3所示，其中 $block\_size$ 是用户定义的参数，用于调整 Block 大小。

$$\begin{aligned} block\_mem = 2 \times num\_layers \times hidden\_size \\ \times block\_size \times sizeof(float16) \end{aligned} \tag{3}$$

因此，假设一个用户请求的长度为 n，占用 GPU block 的数量为 $block\_num$，则其 KV Cache 占用的总内存空间如公式4所示。

$$\begin{aligned} KVCache &= block\_mem \times block\_num \\ &= block\_mem \times \lceil \frac{n}{block\_size} \rceil \end{aligned} \tag{4}$$

由此可以计算出张量交换引入的额外开销。在上述公式中，换入换出传输带宽是由实验环境所决定的，在传输数据量较大时基本保持稳定。而 $block\_size$ 与 $block\_mem$ 在推理任务中均保持不变。因此对于不同的用户请求，其区别仅在于序列长度 n 的不同。

3.4. 内存优化决策器

当 GPU 内存不足时，需要调用内存优化策略。AdaptiveLLM 中的内存优化策略分为张量交换和张量重算两种。根据上文的分析，张量交换引入的额外开销等于 KV Cache 的换出开销与换入开销之和；张量重算引入的额外开销等于 prefill 过程的开销。

Algorithm 1 Mem_Schedule

Input: 运行队列 r, 重算兼等待队列 w, 交换队列 s
Output: 无

1: $sorted(r, key =< priority >, order = asc)$
2: **while** $require\_mem(r) > avail\_gpu\_mem()$ **do**
3: $req \leftarrow r.pop()$ // 优先级最低的用户请求
4: $recomp\_time \leftarrow GET\_RECOMP\_TIME(req)$
5: $swap\_time \leftarrow GET\_SWAP\_TIME(req)$
6: **if** $swap\_time < recomp\_time \wedge$ $kv\_cache\_mem(req) \leq avail\_cpu\_mem()$ **then**
7: $SWAP(req)$ // 交付张量交换执行器
8: $s.append(req)$ // 进入交换队列
9: **else**
10: $RECOMP(req)$ // 交付张量重算执行器
11: $w.append(req)$ // 进入重算兼等待队列
12: **end if**
13: **end while**

张量交换和张量重算带来的额外开销成为阻拦用户请求并发度进一步提升的瓶颈，因此内存优化方式的选择尤为重要。在不同的运行环境中，应该使用不同的内存优化策略，减少额外开销。然而，vLLM 在内存优化策略的选择上并未考虑开销问题。针对使用贪心采样策略的用户请求，其执行张量重算。针对使用并行采样或束搜索采样策略的用户请求，其执行张量交换。因此在面对 GPU 内存瓶颈时难以有效地压缩开销，进而无法提升吞吐率。AdaptiveLLM 则对两种内存优化方式的开销进行比较，选择更优者执行。内存优化决策器的工作流程如算法1所示。

当剩余的 GPU 内存空间不足以存放运行队列在下一次迭代中产生的 KV Cache 时（第 2 行），内存优化决策器进入工作状态。选择运行队列中优先级最低的用户请求（第 3 行），调用张量交换开销分析器和张量重算开销分析器来预测其张量交换和张量重算开销（第 4、5 行）。如果交换开销小于重算开销，则该请求进入交换队列，并交付交换执行器处理（第 6-8 行）；否则进入重算队列，并交付重算执行器处理（第 9-11 行）。以上过程循环执行，直至运行队列在下一次迭代中产生的 KV Cache 能够全部存放到 GPU 内存中。此外，当 CPU 内存不足时，内存优化决策器将直接调用张量重算技术（第 6 行）。

3.5. 用户请求调度器

AdaptiveLLM 维护三个用户请求队列：$waiting$ 队列、$running$ 队列与 $swapped$ 队列。$waiting$ 队列存储初次进入调度系统，还未执行过，或者因张量重算而失去 KV Cache 的用户请求；$running$ 队列存储正在运行（执行 decode 阶段）的用户请求；$swapped$ 队列存储被换出到 CPU 中的用户请求。这三个队列之间拥有以下调度规则：

- $running$ 队列中的用户请求运行完毕后会返回客户端，否则继续运行。
- 当 GPU 内存空余时，$swapped$ 队列中的用户请求可以直接转移至 $running$ 队列中。
- 当 GPU 内存空余时，$waiting$ 队列中的用户请求可以在执行 prefill 阶段后转移至 $running$ 队列。

如果剩余的 GPU 内存空间不足以存储 $running$ 队列在下一次迭代中产生的 KV Cache，则需要内存优化决策器进行抢占调度。如果剩余的 GPU 空间足够，则执行启动调度，以扩充 $running$ 队列，避免浪费 GPU 资源。用户请求调度器将部分请求从 $swapped$ 队列或 $waiting$ 队列中转移至 $running$ 队列中。但由于两种转移方式存在较大差别（是否需要执行 prefill 阶段），因此每次扩充 $running$ 队列时，或者仅从 $swapped$ 队列进行调度，或者仅从 $waiting$ 队列进行调度，而无法同时调度两个队列。用户请求调度器的工作流程如算法2所示。

Algorithm 2 Req_Schedule

Input: 大模型 LLM, 待执行的用户请求队列 L
Output: 无

1: $w \leftarrow L$ // 初始化 waiting 队列
2: $r \leftarrow empty\_list$ // 初始化 running 队列
3: $s \leftarrow empty\_list$ // 初始化 swapped 队列
4: **while** $\neg(w.is\_empty() \wedge s.is\_empty() \wedge r.is\_empty())$ **do**
5: $MemSchedule(r, w, s)$ // (内存不足时) 抢占调度
6: $s\_sche \leftarrow SWAP\_IN\_SCHE()$ // 换入队列构建
7: $w\_sche \leftarrow RECOMP\_SCHE()$ // 重算队列构建
8: **if** $GET\_PRI(w\_sche) \leq GET\_PRI(s\_sche)$ **then**
9: $r = r + s\_sche$ // 换入
10: $s = s - s\_sche$
11: **else**
12: $LLM.PREFILL(w\_sche)$ // 重算
13: $r = r + w\_sche$
14: $w = w - w\_sche$
15: **continue**
16: **end if**
17: $LLM.DECODE(r)$ // 单次推理迭代
18: $r \leftarrow [req \in r | \neg req.is\_finished()]$ // 移除完成的请求
19: **end while**

客户端发送的用户请求进入 *waiting* 队列中，而 *running* 队列和 *swapped* 队列最初为空（第 1-3 行）。当 GPU 内存不足时，调用内存优化算法进行抢占调度（第 5 行），否则执行启动调度。

用户请求调度器尽可能多地寻找能从 *swapped* 队列转移至 *running* 队列的用户请求（第 6 行），和能从 *waiting* 队列转移至 *running* 队列的用户请求（第 7 行）。对它们进行优先级比较（第 8 行），若前者的优先级均值较高，则将其直接转移到 *running* 队列中（第 9-10 行）；若后者的优先级均值较高，则其执行 prefill 阶段后转移至 *running* 队列中，同时直接进入下一轮迭代（第 12-15 行）。需要注意的是，当 GPU 内存不足时，无法实现从 *swapped* 队列或 *waiting* 队列向 *running* 队列的调度，即 *w_sche* 和 *s_sche* 队列均为空，也就不存在后续的优先级比较过程了。

在以上调度操作完成后，*running* 队列应当为非空的，否则推理过程无法继续。*running* 队列执行 decode 阶段作为本次迭代（第 17 行），将已完成的用户请求移除后进入下一次迭代（第 18 行）。

对于一个用户请求，定义其优先级等于处理时间除以序列长度，其中处理时间等于当前时刻减去该用户请求初次进入 *waiting* 队列的时刻。定义用户请求队列的优先级等于所有用户请求优先级的平均值。当用户请求初次进入 *waiting* 队列时，其序列长度较短，优先级增长较为迅速，能够被很快处理。而在等待过程中，其优先级在不断提升，避免了饥饿现象。

4. 实验验证

本章介绍实验部分。第一节为实验平台软硬件配置。第二节介绍 LLM 模型与数据集选取，以及实验参数设置。第三节针对基于开销感知的内存优化策略，进行吞吐率测试。第四节针对基于公平性的用户请求调度策略，进行实时性测试。第五节分析张量交换与张量重算预测误差。第六节进行其它测试工作。

4.1. 实验环境

本文开展实验使用的服务器软硬件配置如表1所示。使用 Intel(R) Xeon(R) CPU 和 NVIDIA A800 80GB GPU 作为硬件环境，使用 CUDA-11.8、PyTorch-2.0.1、Ray2.7.1 以及 vLLM-0.2.5 作为底层框架进行开发。服务器使用 PCIe 连接实现 GPU-CPU 通信。

表 1 实验平台的软硬件配置

软件/硬件	型号/版本
CPU	Intel(R) Xeon(R) CPU @ 2.60GHz
GPU	NVIDIA A800 PCIE 80GB
OS	CentOS Linux 7 (Core)
CUDA	11.8
PyTorch	2.0.1
Ray	2.7.1
vLLM	0.2.5

4.2. 模型与数据集

本文选用 OPT [19]（OPT-13B、OPT-30B）和 Llama [20]（Llama-13B、Llama-32.5B）作为实验模型，在三个常见数据集（Chatbot [21]、Alpaca [22]、Summary [23]）上进行测试。数据集信息如表2所示。

表 2 实验数据集选取

数据集	样本总数	平均输入长度	任务类型
Chatbot	258064	17.02	对话类
Alpaca	68912	19.66	指令类
Summary	1799	340.48	摘要类

Chatbot 和 Alpaca 中大多数序列长度较短，而 Summary 中序列长度展现出很大差异性，且包含长序列。它们涵盖了 LLM 应用程序面临的大部分场景。

实验过程中的参数设置模拟 LLM 应用程序在多任务并发场景下的运行状态。可用 GPU 内存越多，调度器支持的批处理大小就越大，因此推理任务的吞吐率较高。本文将 GPU Block 数量设置为 128，模拟 GPU 内存不足时的用户请求抢占场景。同时，可用 CPU 内存较多时，内存优化决策器偏向于使用张量

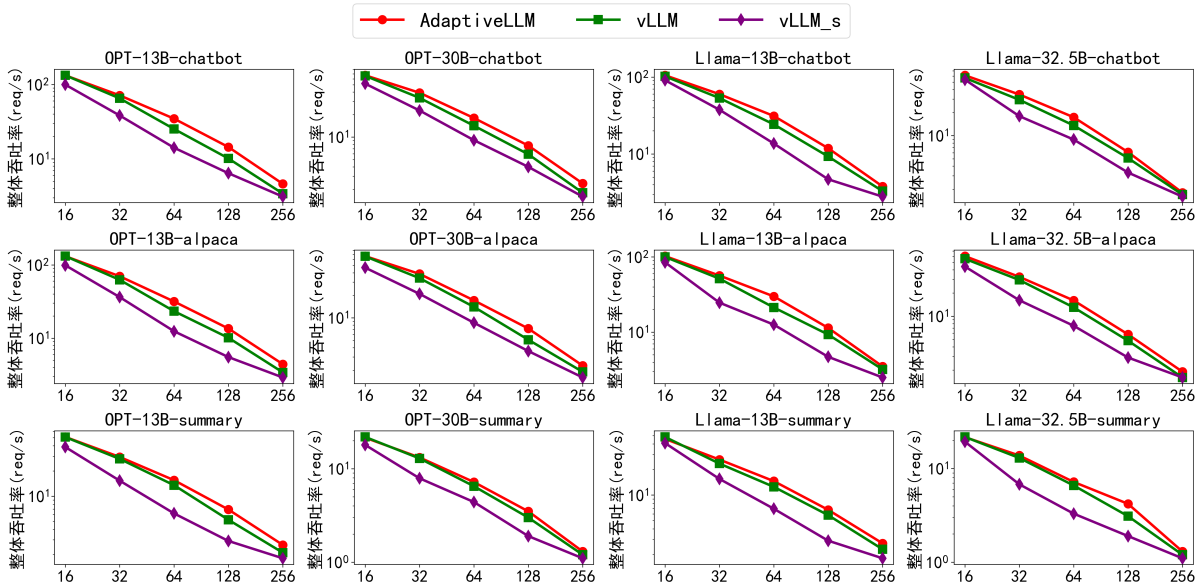

图 3 推理任务吞吐率

交换来释放内存。本文将 CPU Block 数量设置为 64，使得内存优化决策器在 CPU 内存不足时调用张量重算，而非简答地将被抢占用户请求交付张量交换执行器。针对 12 个实验组，在相应数据集中使用简单随机抽样法选取 1000 个样本进行后续测试。

4.3. 吞吐率测试

本文以 vLLM 作为基准框架，针对 AdaptiveLLM 进行吞吐率测试。由于本文在推理过程中使用贪心采样策略来生成新 token，因此 vLLM 内存管理器在 GPU 内存不足时固定调用张量重算技术。同时，对 vLLM 框架稍加修改形成 vLLM_s，使其固定调用张量交换技术。图3展示了 12 个实验组在推理任务中的整体吞吐率测试结果，其横坐标为序列最大输出长度。表3给出了最大输出长度为 64 时，AdaptiveLLM 相对于 vLLM 和 vLLM_s 的具体加速比。结果表明，相比于 vLLM 和 vLLM_s 基准框架，AdaptiveLLM 实现了最高 1.40× 和 2.55× 的整体吞吐加速。

表 3 AdaptiveLLM 相对于基准框架的加速比

LLM-数据集	Chatbot	Alpaca	Summary
OPT-13B	1.38/2.46	1.36/2.55	1.15/2.50
OPT-30B	1.27/1.98	1.22/1.99	1.11/1.64
Llama-13B	1.28/2.28	1.40/2.37	1.17/2.12
Llama-32.5B	1.28/1.95	1.23/2.13	1.09/2.18

由于 Summary 数据集的平均序列长度和方差均明显高于 Alpaca 和 Chatbot 数据集，因此在相同条件下，其推理吞吐率低于 Alpaca 和 Chatbot。尽管如此，与 vLLM 相比，AdaptiveLLM 在 Summary 数据集上也达到了 1.1× 加速比。

表 4 推理任务抢占行为记录

实验组	AdaptiveLLM		vLLM	vLLM_s
抢占行为（千次）	重算	交换	重算	交换
OPT-13B-chatbot	0.11	1.13	1.77	0.78
OPT-13B-alpaca	0.10	1.17	1.82	0.99
OPT-13B-summary	0.10	0.56	0.58	0.26
OPT-30B-chatbot	0.08	1.10	1.64	0.68
OPT-30B-alpaca	0.10	1.05	1.61	0.59
OPT-30B-summary	0.09	0.43	0.47	0.32
Llama-13B-chatbot	0.12	1.02	1.57	0.83
Llama-13B-alpaca	0.08	1.03	1.55	0.87
Llama-13B-summary	0.15	0.55	0.57	0.36
Llama-32.5B-chatbot	0.07	1.04	1.53	0.20
Llama-32.5B-alpaca	0.10	1.00	1.57	0.55

表4给出了序列最大输出长度为 64 时，不同框架推理过程中的抢占行为次数。由表可知，AdaptiveLLM 可以根据模型配置，灵活地选择合适的内存优化策略。CPU 内存的限制使 vLLM_s 的批处理大小低于 vLLM 和 AdaptiveLLM，因此吞吐率较低。当序列最大输出长度限制在较低水平时，每个请求执行推理任务所需的迭代次数较少，资源需求量低，抢占鲜有发生，此时 AdaptiveLLM 和 vLLM 的性能差距不大。随着最大输出长度的增加，有限的 GPU 内存无法满足需求，AdaptiveLLM 调用基于开销感知的内存优化策略，展现性能优势。当最大输出长度过大时，无论是 AdaptiveLLM 还是 vLLM，其批处理大小均限制

在较低水平，但 AdaptiveLLM 仍具有明显优势（当序列最大输出长度为 256 时，AdaptiveLLM 在 vLLM 的基础上实现最高 1.3× 的加速比）。

在 Chatbot 和 Alpaca 数据集的推理任务中，序列长度较短，批处理大小高。根据预测器给出的结果可知，此时张量交换开销小于张量重算开销。然而随着新 token 的不断生成，大量用户请求需要被换出，导致 CPU 内存不足。因此 AdaptiveLLM 执行了大量张量交换操作和少量张量重算操作。

在 Summary 数据集的推理任务中，其序列长度较大，批处理大小低。张量交换发生频率较低，极少出现 CPU 内存不足的现象，此时张量重算操作的执行大部分来源于开销比较的结果。

综上所述，在基于开销感知的内存优化策略下，AdaptiveLLM 在 GPU 内存不足时预测张量交换与张量重算开销，并选择开销较小的内存优化技术执行，进而大幅度提升推理任务整体吞吐率。

4.4. 实时性测试

为了消除整体吞吐率变化对实时性测试的影响，本文选取平均带权周转时间作为测试指标。用户请求带权周转时间等于客户端响应时间除以服务器端处理时间，如公式5所示。对于某一用户请求来说，$finish\_t$ 是其处理完毕时刻，$send\_t$ 是其从客户端发送至服务器端的时刻，$sche\_t$ 是其被 AdaptiveLLM 初次调度的时刻。平均带权周转时间（≥ 1）越低，说明用户请求处理过程中的排队时间占比越低。

$$w\_around\_t = \frac{finish\_t - send\_t}{finish\_t - sche\_t} \tag{5}$$

图4展示了平均带权周转时间随批处理大小上限的变化情况。在不同批处理大小设置下，基于公平性的用户请求调度策略均能使平均带权周转时间显著下降。当批处理大小较大时（64 或 128），AdaptiveLLM 的平均带权周转时间相比于 vLLM 下降了 20% 至 40%，相比于 vLLM_s 下降了 20% 至 60%。

对于序列较短的 Chatbot 和 Alpaca 数据集而言，随着批处理大小的上升，GPU 利用更加充分，因此平均带权周转时间下降。在实际运行中，当批处理大小到达 64 至 128 时，GPU 产生内存瓶颈，此时批处理大小无法继续提升，平均带权周转时间达到最小值。

对于序列较长的 Summary 数据集而言，其处理并发度被限制在较低水平（10 以下），无法达到用户设置的批处理大小上限。因此平均带权周转时间呈稳定状态。AdaptiveLLM 中高效的调度策略展现优势，使用户请求等待时间显著低于 vLLM 和 vLLM_s。

综上所述，基于公平性的用户请求调度策略使得用户请求从客户端发送至服务器端后能够很快开始处理，不会出现长时间等待现象。

4.5. 预测误差测试

(1) 张量重算预测误差

张量重算开销由张量重算开销分析器根据 LLM 层数、LLM 隐藏维度、单请求需要处理的 token 数量、以及批处理大小预测得到。表5和表6分别展示了 OPT 模型和 Llama 模型单步推理执行时间的预测效果。OPT 执行时间预测任务共有 6.4w 条训练数据和 1.6w 条测试数据，结果表明，随机森林回归模型性能最佳，其在拟合 2 次多项式时能够达到 1.76% 的预测误差。Llama 执行时间预测任务共有 6.8 条训练数据和 1.7w 条测试数据，结果表明，随机森林模型同样性能最佳，其在拟合 2 次多项式时能够达到 1.30% 的预测误差。

表 5 OPT 模型单步迭代执行时间预测误差

模型-拟合次数	1	2	3	4	5
线性回归模型	46.52	46.65	28.75	11.86	9.32
决策树	1.81	1.81	1.81	1.81	1.81
随机森林	1.77	1.76	1.77	1.77	1.78
岭回归模型	46.52	46.37	28.45	11.51	7.36
lasso 回归模型	40.22	25.53	27.38	26.08	25.49
弹性回归模型	111.89	123.62	91.67	87.59	86.48
梯度提升模型	15.57	16.05	14.80	15.09	14.68
KNN 回归模型	2.55	2.80	2.89	3.00	3.05

表 6 LLama 模型单步迭代执行时间预测误差

模型-拟合次数	1	2	3	4	5
线性回归模型	76.41	69.44	39.61	12.91	9.18
决策树	1.33	1.32	1.33	1.33	1.34
随机森林	1.31	1.30	1.31	1.31	1.31
岭回归模型	76.41	69.01	39.18	12.73	7.72
lasso 回归模型	69.23	33.57	34.42	35.16	31.58
弹性回归模型	127.18	139.7	100.18	94.94	93.51
梯度提升模型	22.42	21.97	19.42	19.99	19.38
KNN 回归模型	2.24	2.36	2.48	2.63	2.68

(2) 张量交换预测误差

张量交换预测开销由张量交换开销分析器根据用户请求的 KV Cache 内存占用和 GPU-CPU 双向传输带宽而计算得到。本文针对模型 Llama-13B 和 Llama-32.5B 进行测试，其结果如图5所示。两个模型换入开销预测的 MAPE 误差分别为 1.5% 和 1.1%，换

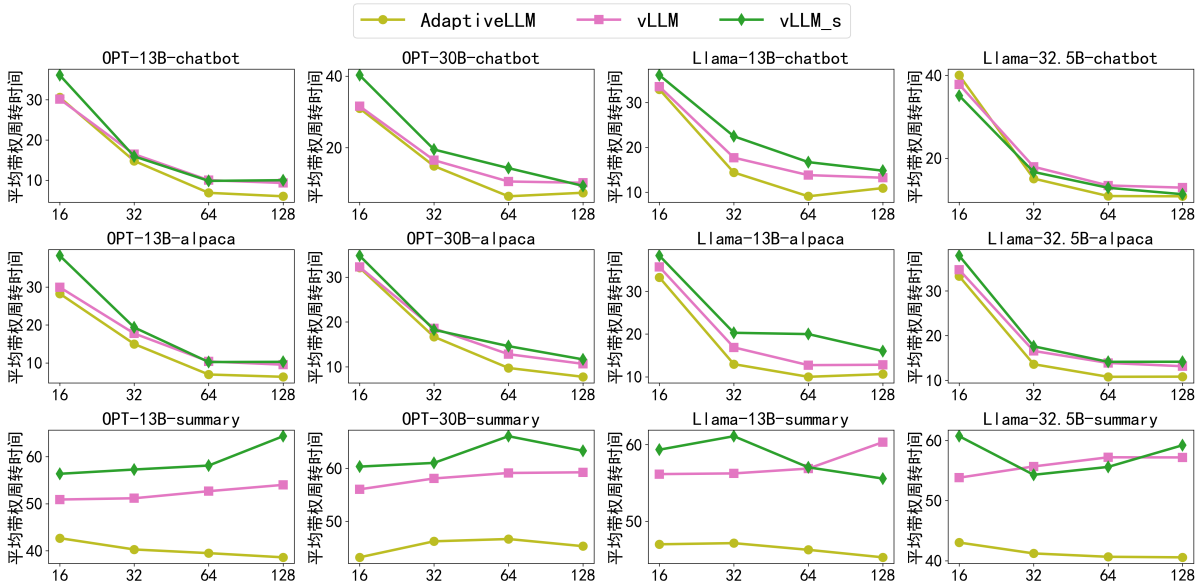

图 4 用户请求平均带权周转时间

出开销预测的 MAPE 误差分别为 1.0% 和 1.2%。因此，张量交换开销总预测误差低于 4%。

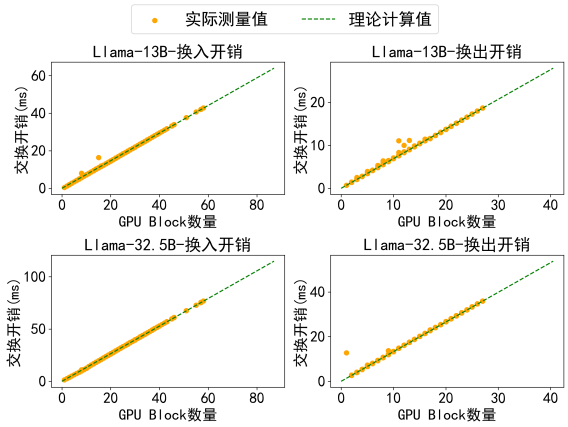

图 5 交换开销预测误差

4.6. 开销测试

基于开销感知的内存优化策略在获取张量重算和张量交换开销时，会带来新的预测开销。本文设计如下对照实验获取预测过程的开销：在吞吐率测试过程中，当 GPU 内存不足时调用开销比较过程，但最终使用 vLLM 提供的固定式内存优化策略（张量重算）。观察此情景下推理任务的总用时可知，预测开销在推理任务中仅占 0.1% 至 1%。

基于公平性的用户请求调度策略也有一定的开销。理论上，该算法的时间复杂度为 $O(r^2)$，其中 r 为 running 队列中用户请求的数量。在实际运行过程中，每次调度的开销在 0.2ms 左右，总开销在推理任务中仅占 0.5% 至 1%。综上所述，AdaptiveLLM 引入的两种优化策略所带来的额外开销均可忽略不计。

5. 相关工作

(1) 张量交换技术

随着批处理大小的增加或模型参数量的扩展，运行时需要保存的张量会超出 GPU 的内存限制。张量交换技术在 GPU 空间不足时开启，将一部分需要保存，而暂时用不到的张量换出至 CPU 中，在计算需要时重新换入 GPU 中。

HuggingFace Accelerate [28] 实现了张量交换技术，但换出与换入的张量仅限于 LLM 的参数张量。LightLLM [27] 能够针对 KV Cache 进行张量交换，但换出的比例设计为定值，无法根据运行时信息调整。

FlexGen [14] 首次提出了“自适应内存优化”的概念，通过线性规划建模在交换方案的可行域内进行搜索，在给定的时间内找到较优解。然而，FlexGen 假设运行队列中的所有用户请求拥有相同的输出长度。在实际情况下，输出长度具有很大的差异性，使得相关理论无法推广。

本文针对 KV Cache 实现张量交换技术，并进行细粒度内存占用建模分析。根据 GPU 内存使用水平，进行实时换出换入调整。

(2) 张量重算技术

Mimose [15, 29-30] 等工作提出了张量重算技术。在抢占式用户请求调度系统中，当某个请求获得执行权时，会检查之前的计算结果是否保存在 GPU 中，如果不在，则需要重新获取这部分计算结果。此时可

以无需将之前存储的计算结果（如果有）从 CPU 或磁盘中换入到 GPU 中，而仅仅对它们进行重新计算。

对于执行 LLM 推理任务的用户请求而言，这些 key-value 张量在初次生成时经历了多次前向传播，而在重算过程中仅需调用自注意力机制即可得到，因此张量重算的开销远远小于 token 序列初次生成时的开销，不会导致计算量的爆炸式增长。

张量交换开销通过简单的计算即可得到，而张量重算开销的计算则略微复杂。Capuchin [31] 将张量重算开销计算过程分解到算子粒度。对于每个算子，通过记录其输入张量与输出张量的生成时间，来获取该算子的重算开销。AdaPipe [32] 将连续出现的多个算子组合成计算单元，通过模拟运行来记录各个计算单元的重算开销。

(3) LLM 推理优化技术

除了张量交换和张量重算等针对张量层面的优化策略以外，传统 LLM 推理框架还采用了很多其他推理优化技术。

ORCA [17] 将批处理调度的粒度从单个用户请求细化为单次推理迭代，化解了用户请求相互等待的性能瓶颈。vLLM [16] 基于 OS 页式内存管理思想，在 ORCA 的基础上引入 Paged Attention 机制。vLLM 相比于 OCRA，大幅度提升显存利用率，增加批处理大小上限，进而提升推理任务的整体吞吐率。

SpecInfer [24] 引入了投机推理技术（Speculative Sampling），根据小型 LLM 的输出来预测大型 LLM 的输出，在大幅度提升推理吞吐率的同时保障了输出质量。DistillSpec [33] 在 SpecInfer 的基础上实现了知识蒸馏技术（Knowledge Distillation，KD），使得输出预测的准确率显著提升。

本文提出的 LLM 推理优化策略能够与调度粒度细化、投机推理等研究工作相兼容。本文设计了规范且友好的用户接口，开发者能够根据任务需求来定义各种参数，极大地方便了有关推理优化的深入研究。

6. 未来工作

(1) 面向截止时间点的调度

用户向服务器端提交推理请求，往往期望在特定时间截点（DDL）前得到完整输出。DDL 离当前时间越近，该任务的紧迫程度就越高，因此在调度时应赋予更高的优先级。对于在 DDL 前确定无法完成的请求，应直接返回客户端，避免浪费服务器资源。

(2) 内存优化与前向传播的并行

- **张量交换与前向传播的并行。**张量交换的本质是 GPU-CPU 通信传输过程，而前向传播的本质是 GPU 计算过程。二者在传统模式下串行执行。AdaptiveLLM 计划在内存优化决策器中设计一个交换线程和一个计算线程，并行完成两项任务，进一步减少张量交换带来的额外开销。

- **张量重算与前向传播的并行。**SARATHI [25] 框架研发了 chunk-prefill 技术，实现 prefill 阶段与 decode 阶段共置运行。由于张量重算的本质是 prefill 过程，因此若将该技术移植到 AdaptiveLLM 中，可以实现张量重算与前向传播的并行。

(3) 张量并行与流水线并行 [34]

AdaptiveLLM 的优化技术仅应用于单个 GPU，未来将扩展至张量并行（Tensor Parallelism，TP）与流水线并行（Pipeline Parallelism，PP）模式。在 TP 场景下，交换与重算开销的计算方式发生变化，模型隐藏维度被均分至不同 GPU 中。在 PP 场景下，交换与重算开销的计算方式与单节点完全相同。由于不同流水线阶段的推理时间无法保证完全相同，因此会无法避免地产生流水线气泡 [32, 35-37]。AdaptiveLLM 希望使用张量交换和张量重算进行气泡填充。

(4) 可扩展性

AdaptiveLLM 所提出的内存优化策略与用户请求调度策略能够应用于不同配置的服务器硬件平台。迁移至新平台后，需要收集张量重算开销，训练单步推理时间预测模型，并获取 GPU-CPU 双向传输带宽信息。整个预处理过程通过数据收集脚本实现，运行时间不超过 10min。本文在未来将针对预处理过程进行优化，以实现轻量级代码迁移。

7. 结论

本文设计了 AdaptiveLLM，一款基于张量交换和张量重算的 LLM 推理服务框架。AdaptiveLLM 实现了张量重算开销预测与张量交换开销预测，其预测误差分别在 2% 和 4% 以下。AdaptiveLLM 研发了基于开销感知的内存优化策略和基于公平性的用户请求调度策略。基于开销感知的内存优化策略在 GPU 内存不足时，执行开销较小的内存优化方式来保证推理任务的顺利完成。基于公平性的用户请求调度策略则能够在 GPU 内存充足时调度更多的用户请求。实验表明，以 vLLM 框架作为基准程序时，AdaptiveLLM 有 10%-40% 的整体吞吐率提升，实现面向服务器端的处理加速。同时能够以合理的方式调度用户请求，将平均带权周转时间降低 20%-40%，实现面向客户端的实时处理。综上所述，AdaptiveLLM 权衡整体吞吐率与单请求延时，化解二者在优化实现上的矛盾。

参 考 文 献

[1] HIDAYAT F, ELVIANI U, SITUMORANG G B G, et al. Face recognition for automatic border control: A systematic literature review[J/OL]. IEEE Access, 2024, 12: 37288-37309. https://doi.org/10.1109/ACCESS.2024.3373264.

[2] DU J. Enhancing personalized recommendations with transferable user representation learning in limited data contexts[D/OL]. University of New South Wales, Sydney, Australia, 2024. http://hdl.handle.net/1959.4/101849. DOI: 10.26190/UNSWORKS/25552.

[3] RAHMAN M M, GUPTA D, BHATT S, et al. A comprehensive review of machine learning approaches for anomaly detection in smart homes: Experimental analysis and future directions[J/OL]. Future Internet, 2024, 16(4): 139. https://doi.org/10.3390/fi16040139. DOI: 10.3390/FI16040139.

[4] GUPTA A, GUPTA A, RAJ G, et al. Traffic light detection for self-driving cars using the yolov8 architecture[C/OL]//Proceedings of the Cognitive Models and Artificial Intelligence Conference, AICCONF 2024, İstanbul, Turkiye, May 25-26, 2024. ACM, 2024: 263-269. https://doi.org/10.1145/3660853.3660925.

[5] BISHOP C M, BISHOP H. Deep learning - foundations and concepts [M/OL]. Springer, 2024. https://doi.org/10.1007/978-3-031-45468-4.

[6] JIM J R, TALUKDER M A R, MALAKAR P, et al. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review [J/OL]. Nat. Lang. Process. J., 2024, 6: 100059. https://doi.org/10.1016/j.nlp.2024.100059. DOI: 10.1016/J.NLP.2024.100059.

[7] CHANG Y, WANG X, WANG J, et al. A Survey on Evaluation of Large Language Models[J/OL]. ACM Trans. Intell. Syst. Technol., 2024, 15(3): 39:1-39:45. https://doi.org/10.1145/3641289.

[8] GOYAL T, LI J J, DURRETT G. News summarization and evaluation in the era of GPT-3[J/OL]. CoRR, 2022, abs/2209.12356. https://doi.org/10.48550/arXiv.2209.12356. DOI: 10.48550/ARXIV.2209.12356.

[9] FENG Z, ZHANG Y, LI H, et al. Improving llm-based machine translation with systematic self-correction[J/OL]. CoRR, 2024, abs/2402.16379. https://doi.org/10.48550/arXiv.2402.16379. DOI: 10.48550/ARXIV.2402.16379.

[10] OUYANG S, ZHANG J M, HARMAN M, et al. LLM is like a box of chocolates: the non-determinism of chatgpt in code generation[J/OL]. CoRR, 2023, abs/2308.02828. https://doi.org/10.48550/arXiv.2308.02828. DOI: 10.48550/ARXIV.2308.02828.

[11] SAITO K, SOHN K, LEE C, et al. Unsupervised LLM adaptation for question answering[J/OL]. CoRR, 2024, abs/2402.12170. https://doi.org/10.48550/arXiv.2402.12170. DOI: 10.48550/ARXIV.2402.12170.

[12] DALE R. GPT-3: What's it good for?[J/OL]. Nat. Lang. Eng., 2021, 27 (1): 113-118. https://doi.org/10.1017/S1351324920000601.

[13] AMINABADI R Y, RAJBHANDARI S, AWAN A A, et al. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale[C/OL]//WOLF F, SHENDE S, CULHANE C, et al. SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, November 13-18, 2022. IEEE, 2022: 46:1-46:15. https://doi.org/10.1109/SC41404.2022.00051.

[14] SHENG Y, ZHENG L, YUAN B, et al. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU[C/OL]// KRAUSE A, BRUNSKILL E, CHO K, et al. Proceedings of Machine Learning Research: volume 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 31094-31116. https://proceedings.mlr.press/v202/sheng23a.html.

[15] LIAO J, LI M, YANG H, et al. Exploiting Input Tensor Dynamics in Activation Checkpointing for Efficient Training on GPU[C/OL]//IEEE International Parallel and Distributed Processing Symposium, IPDPS 2023, St. Petersburg, FL, USA, May 15-19, 2023. IEEE, 2023: 156-166. https://doi.org/10.1109/IPDPS54959.2023.00025.

[16] KWON W, LI Z, ZHUANG S, et al. Efficient Memory Management for Large Language Model Serving with PagedAttention[C/OL]//FLINN J, SELTZER M I, DRUSCHEL P, et al. Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023. ACM, 2023: 611-626. https://doi.org/10.1145/3600006.3613165.

[17] YU G, JEONG J S, KIM G, et al. Orca: A Distributed Serving System for Transformer-Based Generative Models[C/OL]//AGUILERA M K, WEATHERSPOON H. 16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022. USENIX Association, 2022: 521-538. https://www.usenix.org/conference/osdi22/presentation/yu.

[18] NVIDIA. FasterTransformer[EB/OL]. 2020. https://github.com/NVIDIA/FasterTransformer.

[19] ZHANG S, ROLLER S, GOYAL N, et al. OPT: Open Pre-trained Transformer Language Models[J/OL]. CoRR, 2022, abs/2205.01068. https://doi.org/10.48550/arXiv.2205.01068. DOI: 10.48550/ARXIV.2205.01068.

[20] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and Efficient Foundation Language Models[J/OL]. CoRR, 2023, abs/2302.13971. https://doi.org/10.48550/arXiv.2302.13971. DOI: 10.48550/ARXIV.2302.13971.

[21] PALLA A. Chatbot Instruction Prompts[EB/OL]. 2023. https://huggingface.co/datasets/alespalla/chatbot_instruction_prompts.

[22] BHARTI G. Finance Alpaca[EB/OL]. 2023. https://huggingface.co/datasets/gbharti/finance-alpaca.

[23] ALI K. Summary Dataset[EB/OL]. 2024. https://huggingface.co/datasets/khwrali011/summary-dataset.

[24] MIAO X, OLIARO G, ZHANG Z, et al. SpecInfer: Accelerating Generative LLM Serving with Speculative Inference and Token Tree Verification [J/OL]. CoRR, 2023, abs/2305.09781. https://doi.org/10.48550/arXiv.2305.09781. DOI: 10.48550/ARXIV.2305.09781.

[25] AGRAWAL A, PANWAR A, MOHAN J, et al. SARATHI: Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills[J/OL]. CoRR, 2023, abs/2308.16369. https://doi.org/10.48550/arXiv.2308.16369. DOI: 10.48550/ARXIV.2308.16369.

[26] ZHENG Z, REN X, XUE F, et al. Response Length Perception and Sequence Scheduling: An LLM-Empowered LLM Inference Pipeline [C/OL]//OH A, NAUMANN T, GLOBERSON A, et al. Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023. 2023. http://papers.nips.cc/paper_files/paper/2023/hash/ce7ff3405c782f761fac7f849b41ae9a-Abstract-Conference.html.

[27] GITHUB. lightllm[EB/OL]. 2024. https://github.com/ModelTC/lightllm.

[28] HUGGINGFACE. Huggingface Accelerate[EB/OL]. 2022. https://huggingface.co/docs/accelerate/index.

[29] JAIN P, JAIN A, NRUSIMHA A, et al. Checkmate: Breaking the memory wall with optimal tensor rematerialization[C/OL]//DHILLON I S, PAPAILIOPOULOS D S, SZE V. Proceedings of the Third Conference on Machine Learning and Systems, MLSys 2020, Austin, TX, USA, March 2-4, 2020. mlsys.org, 2020. https://proceedings.mlsys.org/paper_files/paper/2020/hash/0b816ae8f06f8dd3543dc3d9ef196cab-Abstract.html.

[30] CHEN T, XU B, ZHANG C, et al. Training deep nets with sublinear memory cost[J/OL]. CoRR, 2016, abs/1604.06174. http://arxiv.org/abs/1604.06174.

[31] PENG X, SHI X, DAI H, et al. Capuchin: Tensor-based GPU Memory Management for Deep Learning[C/OL]//LARUS J R, CEZE L, STRAUSS K. ASPLOS '20: Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, March 16-20, 2020. ACM, 2020: 891-905. https://doi.org/10.1145/3373376.3378505.

[32] SUN Z, CAO H, WANG Y, et al. AdaPipe: Optimizing Pipeline Parallelism with Adaptive Recomputation and Partitioning[C/OL]//GUPTA R, ABU-GHAZALEH N B, MUSUVATHI M, et al. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024. ACM, 2024: 86-100. https://doi.org/10.1145/3620666.3651359.

[33] ZHOU Y, LYU K, RAWAT A S, et al. Distillspec: Improving speculative decoding via knowledge distillation[C/OL]//The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. https://openreview.net/forum?id=rsY6J3ZaTF.

[34] BRAKEL F, ODYURT U, VARBANESCU A L. Model parallelism on distributed infrastructure: A literature review from theory to LLM case-studies[J/OL]. CoRR, 2024, abs/2403.03699. https://doi.org/10.48550/arXiv.2403.03699. DOI: 10.48550/ARXIV.2403.03699.

[35] HUANG Y, CHENG Y, BAPNA A, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism[C/OL]//WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 103-112. https://proceedings.neurips.cc/paper/2019/hash/093f65e080a295f8076b1c5722a46aa2-Abstract.html.

[36] FAN S, RONG Y, MENG C, et al. DAPPLE: a pipelined data parallel approach for training large models[C/OL]//LEE J, PETRANK E. PPoPP '21: 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Virtual Event, Republic of Korea, February 27- March 3, 2021. ACM, 2021: 431-445. https://doi.org/10.1145/3437801.3441593.

[37] HARLAP A, NARAYANAN D, PHANISHAYEE A, et al. Pipedream: Fast and efficient pipeline parallel DNN training[J/OL]. CoRR, 2018, abs/1806.03377. http://arxiv.org/abs/1806.03377.