AdaptiveLLM: 基于张量交换和张量重算的大模型推理优化技术

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

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 [5]（Deep Learning，DL）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 [6]（Natural Language Processing，NLP）作为深度学习领域的重要研究方向，长期以来备受人们关注。近年来，随着 GPU 算力的不断提升，各种语言模型也朝着复杂化，多功能化的方向迅猛发展。

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，因此需要海量的文本数据进行训练。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，拥有巨大的科研价值与商业价值。2021 年 GPT-3 模型 [8, 12] 的问世标志着 LLM 领域的里程碑，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，使得该领域的热度空前高涨。

复杂的结构和爆炸式增长的参数量为 LLM 带来了卓越的 NLP 性能，但却为众多的科研工作者们带来了巨大难关。LLM 在训练时消耗资源多，花费时间长，失败风险高，性能要求严，使得 LLM 的训练成本急剧上涨。推理任务的成本相对较低，但极高的内存占用也为推理任务的高效执行带来了独特的挑战。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]，在传统的 LLM 推理框架下，需要使用至少 5 个 NVIDIA A100 GPU（80GB），且需要引入复杂的并行化策略。因此，降低运行时资源消耗对 LLM 推理任务至关重要。

复杂的结构和爆炸式增长的参数量为 LLM 带来了卓越的 NLP 性能，但却为众多的科研工作者们带来了巨大难关。LLM 在训练时消耗资源多，花费时间长，失败风险高，性能要求严，使得 LLM 的训练成本急剧上涨。推理任务的成本相对较低，但极高的内存占用也为推理任务的高效执行带来了独特的挑战。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，在传统的 LLM 推理框架下，需要使用至少 5 个 NVIDIA A100 GPU（80GB），且需要引入复杂的并行化策略。

为了应对较高参数量带来的 GPU 内存瓶颈，本文提出了一款基于张量交换 [14]（Swapping）和张量重算 [15]（Recomputation）的 LLM 推理服务框架，引入先进的显存优化策略和用户请求调度策略实现 LLM 的高效推理。该框架针对服务器端实现高吞吐，针对客户端实现实时请求处理，进一步解决吞吐率与单请求延时在性能优化上长期以来存在的矛盾。

传统的 LLM 推理框架拥有诸多可改进之处，下面列举其中最为显著的两点。

首先，通过张量交换和张量重算等内存优化技术，可以在有限的 GPU 内存空间中增加批处理大小。然而，上述两项张量优化技术对 LLM 推理性能的影响十分复杂，取决于服务器硬件环境（如 GPU 计算能力、GPU-CPU 传输带宽）、用户设置（如新 token 的采样方式）、LLM 与数据集选取、以及推理任务的运行时信息等等。已有的推理框架 [14, 16-17] 在面对 GPU 内存瓶颈时固定调用张量交换或张量重算技术，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。

其次、在针对 LLM 推理任务进行性能优化时，传统工作 [14, 16, 18] 或者以整体吞吐率为单一导向，或者以单请求延时为单一导向，而没有在二者间进行权衡。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。二者均在 LLM 应用程序中占有重要地位。

针对传统工作的不足之处，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。具体而言，本文开展了以下工作：

- **本文设计了一款张量重算分析器，实现张量重算开销的精准预测。**

  通过算子粒度的计算复杂度分析来找出张量重算开销的影响因素，而后模拟 LLM 的前向传播，收集数据，并建立单步迭代执行时间的回归预测模型。实验表明，张量重算开销的预测误差在 2% 以内。

- **本文设计了一款张量交换分析器，实现张量交换开销的精准预测。**

  通过获取用户请求 KV Cache 的内存占用和 GPU-CPU 间通信效率，来计算数据传输开销。实验证明，张量交换开销预测误差在 4% 以内。

- **本文设计了一个基于张量交换和张量重算的自适应 LLM 推理加速器。**

  通过引入基于开销感知的张量优化策略，提升整体吞吐率，实现了面向服务器端的处理加速；通过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，实现了面向客户端的实时请求处理。

- **本文搭建了一款 LLM 推理服务框架 AdaptiveLLM，并进行实验评估。**

AdaptiveLLM 实现了上述设计的张量重算分析器、张量交换分析器与自适应 LLM 推理优化器等功能模块。以 vLLM 框架作为基础程序，在典型 LLM（OPT[19]、Llama[20]）和数据集（Summary[21]、Chatbot[22]、Alpaca[23]）上进行实验验证。结果表明，本文能够实现 10% 到 40% 的吞吐率提升，且将用户请求平均带权周转时间降低为 60% 至 80%。以此证明本文的优化技术实现了整体吞吐率与单请求延时的权衡，完成了 LLM 的高效推理。

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 张量进行优化，因此本章将在第一节阐明 KV Cache 在 LLM 推理任务中的功能；第二节论述两个面向 KV Cache 的常见优化方向；第三节列举两个传统工作中针对 KV Cache 的张量优化技术。

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时间开销。主流 LLM 推理服务 [14, 16-18] 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，其工作原理如图1所示。

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 的参数张量，KV Cache 在对应用户请求推理完毕后丢弃，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。

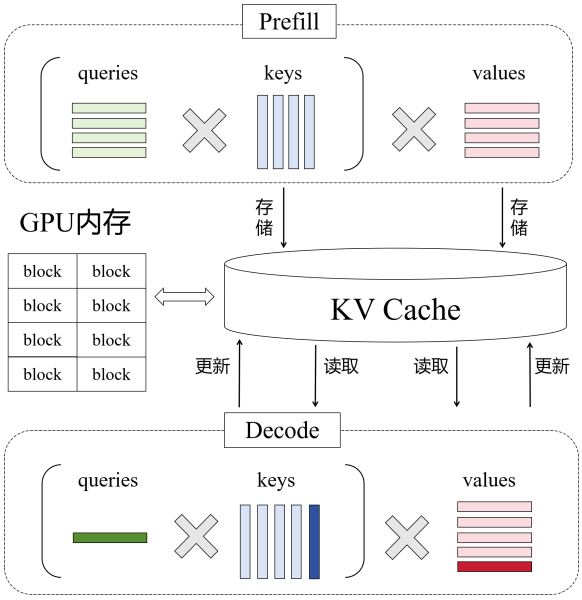

图 1 KV Cache 的功能示意图

2.2. KV Cache 的优化瓶颈

KV Cache 的引入方便了计算过程，却带来一些挑战，使得 LLM 推理性能的提升无法达到预期水平。下面介绍两个被人们所广泛讨论和研究的问题。

(1) 内存利用率问题

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 [24] 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) 通信开销问题

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, 25]，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，内存管理器放弃张量的换出和换入过程，在用户请求被调度时执行一次 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张量优化策略，能够预测二者的开销，并选择开销小的过程执行。

2.3. 针对 KV Cache 的张量优化技术

在 LLM 推理服务过程中，传统的张量优化（也称抢占）技术有三种：张量交换、张量重算和张量压缩 [14]。AdaptiveLLM 实现了张量交换与张量重算。而张量压缩目前还未能实现，将在本文第五章介绍。

(1) 张量交换

服务器拥有 GPU-CPU-磁盘三级存储结构。GPU 位于三级存储中的最上层，其计算速率快，并行度高，但存储空间有限，而 CPU 和磁盘的存储空间相对较大。为了提升服务器的实时吞吐率，LLM 推理服务框架一般采用批处理的方式执行用户请求。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。当检测到 GPU 内存占用峰值达到较高水平时，需要开启张量交换功能，将一部分需要保存，而暂时用不到的张量换出到 CPU 甚至磁盘中，在计算需要时重新换入 GPU 中。综上所述，张量交换包括换出与换入两个阶段，有两次数据传输过程。

(2) 张量重算

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。

3. 优化设计

本章介绍创新工作。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 借鉴了 vLLM 框架的设计思想，在此基础上开发了三个新功能模块，包括张量重算分析器、张量交换分析器和自适应 LLM 推理加速器，其整体架构如图2所示。

张量重算分析器基于用户请求长度、模型隐藏维度、模型层数等运行时信息来预测重算开销；张量交换分析器基于 KV Cache 内存占用和 GPU-CPU 双向传输带宽来预测交换开销。实验环节将会计算二者的预测误差。

自适应 LLM 推理加速器包含内存优化决策器和用户请求调度器。内存优化决策器引入了基于开销感知的张量优化策略，提升总体吞吐率，实现面向服务器端的处理加速；用户请求调度器引入了基于公平性的用户请求调度策略，减少单请求响应时间，实现面向客户端的实时处理。

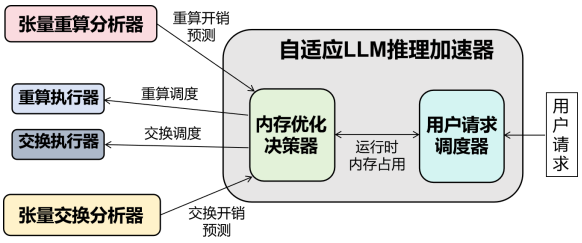

图 2 整体设计架构

在推理过程中，内存优化决策器与用户请求调度器共享运行时内存占用信息。在 GPU 内存不足时，内存优化决策器按照一定的算法选择优先级最低的用户请求进行抢占。其收集张量重算分析器提供的重算开销预测值与张量交换分析器提供的交换开销预测值，选择开销较小的抢占方式，而后交付相应的执行器。用户请求调度器在 GPU 内存空余时重新调度被抢占的用户请求，在满足公平性的前提下尽可能多地调度用户请求，避免 GPU 资源的浪费。二者高效协同，实现整体吞吐率与单请求延时的权衡。

本章第二小节至第五小节将依次介绍张量重算分析器、张量交换分析器、内存优化决策器和用户请求调度器的设计原理与实现细节。

3.2. 张量重算分析器

张量重算技术的时间线流程如图3所示。当用户请求被抢占时，重算执行器在内存中删除其 KV Cache；被重新调度时，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，通过算子粒复杂度分析来定位单步

推理时间的影响因素。

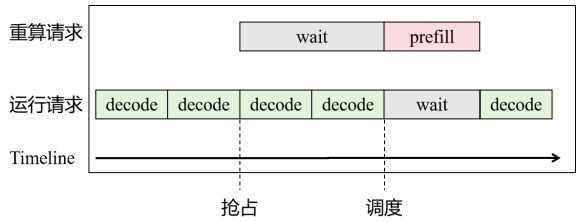

图 3 张量重算时间线流程

(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4所示。图中 X_i,Y_i 是由用户输入决定的张量维度；$input\_dim$,$output\_dim$,$head\_size$ 是由算子决定的张量维度。

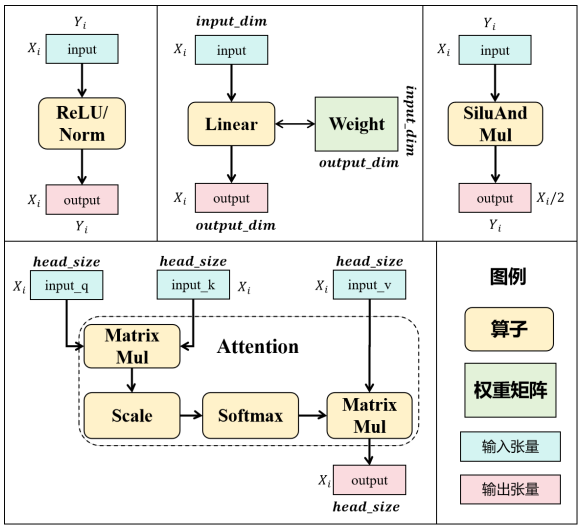

图 4 四种算子的计算流程

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**：是 RowParallelLinear，ColumnParallelLinear 等多种线性层的统称，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，可以找出 4 项有关 LLM 单步推理执行时间的影响因素，分别为：LLM 层数、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 9 个回归模型，包括线性回归模型、支持向量机回归模型、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 5）进行遍历测试。选择在测试集上预测误差最小者及相应的多项式次数，将其部署到 AdaptiveLLM 的张量重算分析器中。

3.3. 张量交换分析器

张量交换的时间线流程如图5所示。当用户请求被抢占时，交换执行器将其 KV Cache 从 GPU 中传输到 CPU 中（换出阶段）；被重新调度时，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换出时间与换入时间之和。

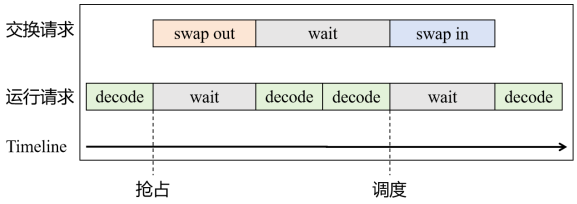

图 5 张量交换时间线流程

换出开销与换入开销的计算方式如公式2所示。

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

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 内存占用的计算公式为3，其中 $block\_size$ 是用户定义的参数，用于调整 Block 大小。

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

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 占用的总内存空间为公式4。

$$\begin{aligned} KVCache &= 2 * block\_mem * block\_num \\ &= 2 * block\_mem * \lceil \frac{n}{block\_size} \rceil \end{aligned} \tag{4}$$

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，需要启动张量优化策略。AdaptiveLLM 中的张量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。

张量交换和张量重算所带来的额外开销成为了阻拦用户请求并发度进一步提升的瓶颈，因此抢占方式的选择尤为重要，在不同的运行环境中，应该使用不同的抢占策略来实现较低的抢占开销。然而，vLLM 在抢占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，其执行重算抢占；针对使用并行采样或束搜索采样策略的用户请求，其执行交换抢占。这种固定式抢占策略使得 vLLM 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。本文则对两种抢占方式的开销进行比较，选择开销小的抢占方式执行。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当前用户请求批次在下一个迭代中产生的 KV Cache 时（第 2 行），内存优化决策器进入工作状态。选择当前批次中优先级最低的用户请求（第 3 行），调用张量交换分析器和张量重算分析器来预测其交换和重算开销（第 4、5 行）。如果交换开销小于重算开销，则将该请求交付交换执行器处理（第 7、8 行）；否则交付重算执行器处理（第 10、11 行）。以上过程循环执行，直至当前批次在下一个迭代中产生的 KV Cache 能够全部存放到 GPU 内存中。

Algorithm 1 Mem_Schedule

Input: 运行队列 $running$, 重算兼等待队列 $waiting$, 交换队列 $swapped$
Output: 无

```
1:  sorted(running, key =< priority >, order = asc)
2:  while require_mem(running) > available_mem() do
3:      req ← running.pop()
4:      recomp_time ← GET_RECOMP_TIME(req)
5:      swap_time ← GET_SWAP_TIME(req)
6:      if swap_time < recomp_time then
7:          SWAP(req)                    // 交付张量交换执行器
8:          swapped.append(req)
9:      else
10:         RECOMP(req)                  // 交付张量重算执行器
11:         waiting.append(req)
12:     end if
13: end while
```

另外，当 CPU 内存不足时，内存优化决策器将直接调用张量重算，而跳过开销预测和比较过程（由于此情况比较复杂，因此在算法中没有体现）。

3.5. 用户请求调度器

本文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 内存条件允许时，$swapped$ 队列中的用户请求可以直接转移至 running 队列中。
- 当 GPU 内存条件允许时，$waiting$ 队列中的用户请求可以转移至 $running$ 队列中，但需要先执行 prefill 阶段。

如果剩余的 GPU 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，则考虑扩充 $running$ 队列，以避免浪费 GPU 资源。在扩充 $running$ 队列时，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 ← L                                          // 初始化 waiting 队列
2: r ← empty_list                                 // 初始化 running 队列
3: s ← empty_list                                 // 初始化 swapped 队列
4: while ¬(w.is_empty() ∧ s.is_empty() ∧ r.is_empty()) do
5:     MemSchedule(r, w, s)                       // (内存不足时) 抢占调度
6:     s_sche ← SWAP_IN_SCHE()                    // 换入队列构建
7:     w_sche ← RECOMP_SCHE()                     // 重算队列构建
8:     if GET_PRI(w_sche) ≤ GET_PRI(s_sche) then
9:         r = r + s_sche                         // 换入调度
10:        s = s − s_sche
11:    else
12:        LLM.PREFILL(w_sche)                    // 重算调度
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:    for req in r do
19:        if req.is_finished() then
20:            r.remove(req)                      // 移除完成的请求
21:        end if
22:    end for
23: 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），否则扩充 $running$ 队列。

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（第 12 行）转移至 $running$ 队列中（第 13-14 行），同时直接进入下一轮迭代（第 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，此时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（第 17 行），将已完成的用户请求移除后进入下一次迭代（第 18-22 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。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，而序列长度指用户输入与已生成 tokens 的总长度。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，因此优先级增长较为迅速，能够被很快处理；而在用户请求等待过程中，其优先级在不断提升，因此避免了饥饿现象。

vLLM 基于 FCFS 策略进行设计，在调度时优先考虑 $swapped$ 队列，只有当 $swapped$ 队列为空时才调度 $waiting$ 队列，使得以交换方式被抢占的用户请求相比于以重算方式被抢占的用户请求，其重调度的优先级更高。结合上一小节关于 vLLM 固定式抢占策略的分析可知，一部分用户请求被抢占后能够很快重新调度，而也有一部分用户请求被抢占后进入 $waiting$ 队列的末位，需要长时间等待。这种调度策略违反了公平性原则。本文中的用户请求调度器基于公平性原则而设计，同时在实验部分证明，其能够大幅提升用户请求的实时性。

4. 实验验证

本章介绍实验部分。第一节为实验平台的软硬件配置；第二节介绍 LLM 与数据集的选取，以及实验组设置；第三节分析单步迭代执行时间预测误差；第四节针对基于开销感知的张量优化策略，进行吞吐率优化测试；第五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。

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

使用 Intel(R) Xeon(R) CPU 和 NVIDIA A800 80GB GPU，CUDA 版本为 11.8。使用 pytorch-2.0.1、ray-2.7.1 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。PCIe 传输带宽在传输数据量不同时差异显著，表2列出了部分情况。在计算张量交换开销时，需要根据单次

实际传输的数据量（一般是一个 Block 或其中的一部分）找到对应的传输带宽值。

表 2 PCIe 双向传输带宽

传输量（B）	HtoD（MB/s）	DtoH（MB/s）
1024	0.19	0.24
2048	0.60	0.72
4096	1.20	1.49
8192	1.07	2.97
16384	4.16	5.79
32768	7.76	9.35
102400	14.27	16.49
204800	18.30	20.24
409600	21.17	22.57

4.2. 实验设置

本文选用 OPT-13B、OPT-30B、Llama-13B 和 Llama-32.5B 作为实验模型，在三个常见数据集上进行测试，如表3所示。

表 3 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

三个数据集的样本序列长度分布曲线如图6所示。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 应用程序面临的绝大部分场景。

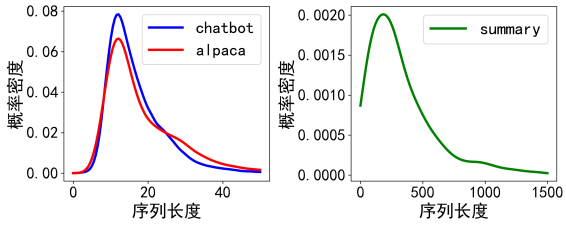

图 6 序列长度分布曲线

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。本文将 GPU Block 数量设置为 128，以保证推理任务执行过程中会发生抢占现象；将 CPU Block 数量设置为 64，以保证张量重算技术能够在 CPU 内存不足时被调用。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. 执行时间预测

表4和表5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 4 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
支持向量机	27.76	23.51	17.88	14.03	11.29
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 5 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
支持向量机	55.82	37.50	26.44	22.92	19.19
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

4.4. 重算与交换的开销对比

本文对 OPT-13B、OPT-30B、Llama-13B 和 Llama-32.5B 进行开销对比测试，其结果如图7所示。当序列长度较小时，交换开销小于重算开销；随着序列长度的增加，二者大小关系反转。原因如下：

自注意力机制内核采用并行计算策略，每个线程只计算一个 token 的 qkv 张量及注意力值。随着 token 数量增多，并行执行的线程数量增加，线程间同步开销随之上升，而单线程计算量不变。因此张量重算开销随序列长度增加呈亚线性增长。而由公式2、3、4可知，张量交换开销与序列长度呈近似正比关系。因此，张量重算开销的增长速度小于张量交换开销。

在贪心采样策略下，对于长序列（如 Summary 数据集中的部分样本），无论是 vLLM 还是 AdaptiveLLM，都偏向于使用重算，两种策略带来的抢占行为没有差异。对于短序列（如 Chatbot 和 Alpaca 数据集），vLLM 使用重算，而 AdaptiveLLM 使用开销较小的交换，此时能够带来吞吐率提升。在 LLM 实际应用场景中，大多数序列的长度较短，使得张量交换在提升性能上拥有明显优势。而当长序列较多，或者 CPU 内存空间不足时，张量重算技术能够发挥优势。

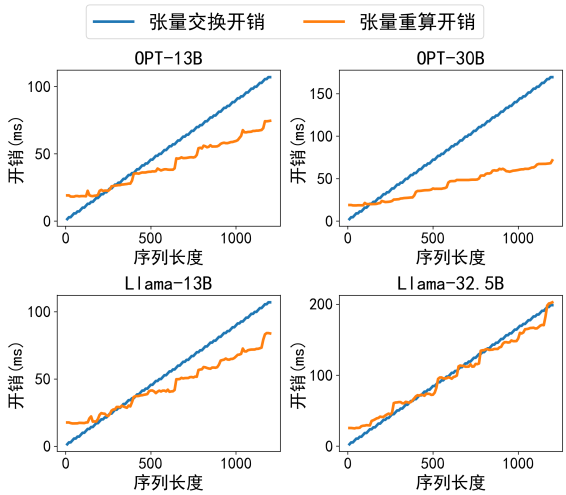

图 7 交换与重算开销对比

4.5. 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。同时，对 vLLM 框架稍加修改形成 vLLM_s，使得内存管理器在 GPU 内存不足时固定调用张量交换技术。

图8展示了 12 个实验组在推理任务中的整体吞吐率与序列最大输出长度的关系。表6给出了最大输出长度为 64 时，AdaptiveLLM 相对于 vLLM 的加速比。结果表明，以 vLLM 为基准框架时，基于开销感知的张量优化策略实现 1.1 到 1.4 的整体吞吐加速比。

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。同理，Alpaca 在相同条件下的推理吞吐率应略低于 Chatbot。

表 6 AdaptiveLLM 相对于 vLLM 的加速比

LLM-数据集	Chatbot	Alpaca	Summary
OPT-13B	1.377	1.356	1.148
OPT-30B	1.268	1.221	1.108
Llama-13B	1.284	1.404	1.168
Llama-32.5B	1.279	1.231	1.091

表7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。

表 7 用户请求抢占行为记录

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

在 GPU 内存不足时，vLLM 调用张量重算技术；vLLM_s 调用张量交换技术；而 AdaptiveLLM 能够从二者中选择开销较小的抢占方式。CPU 内存的限制使 vLLM_s 的批处理大小低于 vLLM 和 AdaptiveLLM，因此吞吐率也较低。当用户请求序列的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张量优化策略，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（当批处理大小为 256 时，AdaptiveLLM 在 vLLM 的基础上实现了 1.1 到 1.3 的整体加速比）。

由表7可知，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，换出频繁，导致 CPU 内存不足，因此 AdaptiveLLM 执行了少量张量重算操作；在 Summary 数据集的推理任务中，其序列长度较大，批处理大小低，换入换出较少，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。

另外，短序列被抢占时，张量交换相比于张量重算有开销优势。而 OPT-13B 和 Llama-13B 相比于 OPT-30B 和 Llama-32.5B，这种优势更加明显，如图7所示。因此在 OPT-13B 和 Llama-13B 上，AdaptiveLLM 产生的加速比（以 vLLM 作为对照）更高。

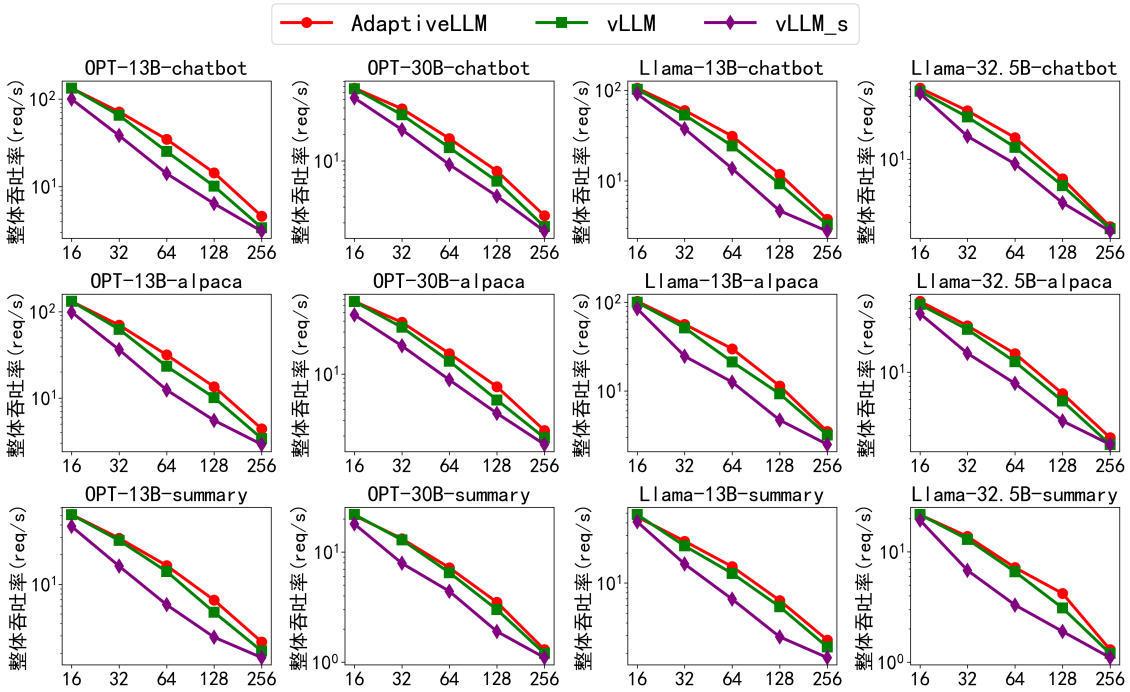

图 8 推理任务吞吐率

4.6. 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。该指标越低，说明用户请求的排队时间越短。

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

图9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为 vLLM 的 60% 至 80%。

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。当批处理大小到达 64 时，GPU 产生内存瓶颈，此时平均带权周转时间不再随最大批处理大小的上升而下降。

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。

4.7. 其它测试

(1) 误差分析

张量重算（单次 prefill 阶段执行）开销的预测误差等于单步推理执行时间预测器的误差，根据本章第三节的分析可知，其预测误差低于 2%。

本文针对模型 Llama-13B 和 Llama-32.5B 进行交换误差测试，其结果如图10所示。两个模型换入开销预测 MAPE 误差分别为 1.5% 和 1.1%；换出开销预测 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%。

(2) 开销分析

基于开销感知的张量优化策略在获取重算和交换开销时，会带来新的预测开销。本文设计如下对照实验获取张量感知过程的开销：

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 提供的固定式张量抢占策略（重算）。观察此情景下推理任务的总用时可知，张量感知过程的开销在整个推理任务中仅占 0.1% 至 1%。

5. 相关工作

传统 LLM 推理框架采取了很多显存优化技术。Hugging Face Accelerate[26] 实现了张量交换技

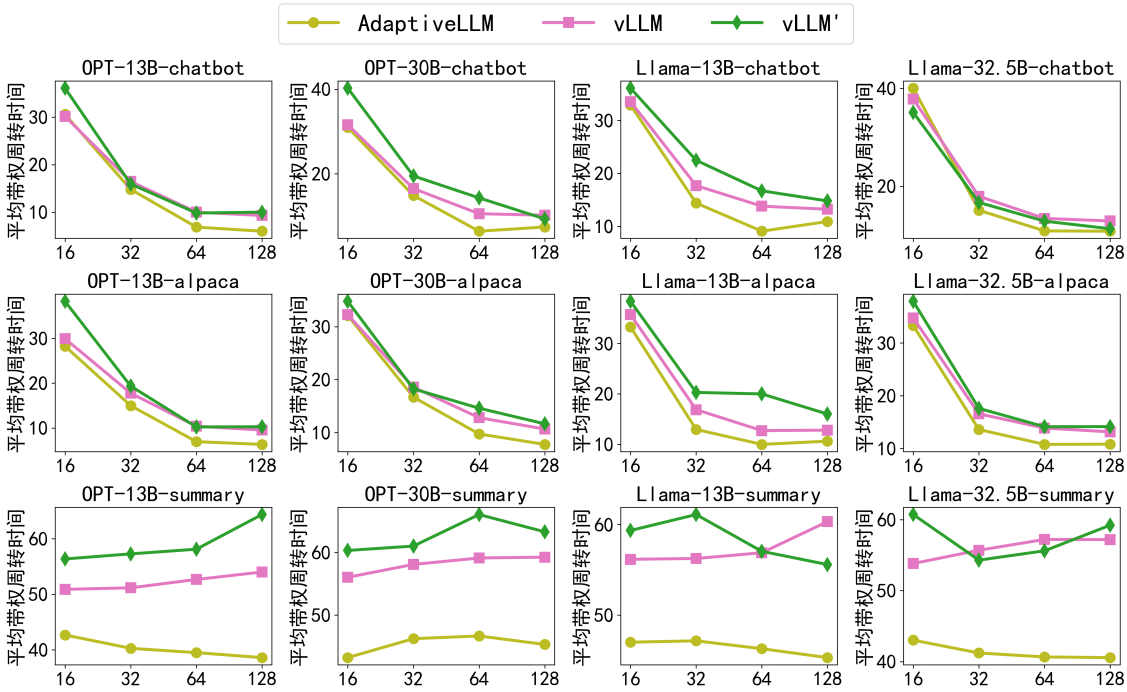

图 9 用户请求平均带权周转时间

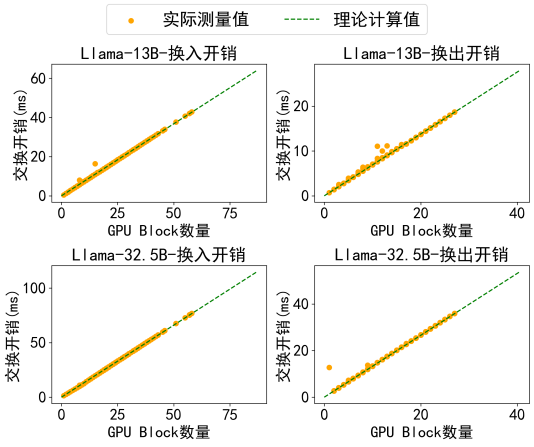

图 10 交换开销预测误差

术，但换出与换入的张量仅限于 LLM 的参数张量；DeepSpeed ZeRO-Inference[13] 实现了 LLM 的分布式推理，能够利用数据并行性与张量并行性来实现 LLM 推理加速。但以上两个框架均无法针对 KV Cache 实现交换技术，也没有实现压缩或重算技术。

FlexGen[14] 首次提出了“自适应张量优化”的概念，并将张量交换的范围由参数张量扩展至所有张量。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。同时，FlexGen 还实现了张量压缩技术，相比于 Hugging Face Accelerate 和 DeepSpeed ZeRO-Inference，实现了较大的吞吐率提升。然而，FlexGen 假设同一批中的所有用户请求拥有相同的输出长度。在实际情况中，输出长度具有很大的差异性，使得相关理论无法推广。

为了解决上述问题，ORCA[17] 将批处理调度的粒度从单个用户请求转化为单次推理迭代，化解了同一批中用户请求相互等待的性能瓶颈。

vLLM[16] 在 ORCA 的基础上实现了张量重算技术。基于 OS 页式内存管理思想，引入 Paged Attention 机制来实现。vLLM 相比于 OCRA，大幅度提升了显存利用率，并增加批处理大小上限，进而提升了推理任务的整体吞吐率。同时，vLLM 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。

另外，其它推理加速技术也被广泛提出和应用。SpecInfer[18] 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[27] 在 SpecInfer 的基础上实现了知识蒸馏技术（Knowledge Distillation，KD），使得输出预测准确率显著提升。

6. 未来工作

在未来一段时间内，本文将针对以下部分完善 AdaptiveLLM 的设计。

(1) 张量压缩技术

张量在三级存储结构中的传输带来无法忽略的通信开销。由于三级存储结构间的传输带宽有限，因

此随着 LLM 参数量和批处理大小的增加，通信开销成为主要性能瓶颈。压缩技术常与交换技术联合使用，通过矩阵变换等数学方式减少传输参数量。高效的压缩技术能够在不损失张量精度的前提下减小通信开销，进一步提升批处理大小上限，提升吞吐率。

(2) 张量优化与前向传播的并行

张量交换：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。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：SARATHI[28] 框架研发了 chunk-prefill 技术，实现 prefill 阶段与 decode 阶段的共置运行。由于张量重算的本质是 Prefill 阶段的执行，因此若将该技术移植到 AdaptiveLLM 中，则能够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 [29]

AdaptiveLLM 目前仅针对张量并行度与流水线并行度均为 1 的场景进行优化，将在未来实现张量并行技术与流水线并行技术。

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张量优化策略和基于公平性的用户请求调度策略。基于开销感知的张量优化策略用于在 GPU 内存不足时，执行开销较小的抢占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 内存充足时重新调度被抢占的用户请求。实验表明，相比于 vLLM 框架，AdaptiveLLM 有 10%-40% 的整体吞吐率提升，实现了服务器端的处理加速；且 AdaptiveLLM 能够以合理的方式调度用户请求，将平均带权周转时间优化为 vLLM 的 60% 80%，减少等待时间，实现了面向客户端的实时请求处理。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] Fadhil Hidayat, Ulva Elviani, George Bryan Gabriel Situmorang, 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] Jing Du. 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] Md. Motiur Rahman, Deepti Gupta, Smriti Bhatt, 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] Aditya Gupta, Ayushi Gupta, Gaurav Raj, 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] Christopher M. Bishop, Hugh Bishop. Deep learning - foundations and concepts[M/OL]. Springer, 2024. https://doi.org/10.1007/978-3-031-45468-4.

[6] Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, 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] Yupeng Chang, Xu Wang, Jindong Wang, 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] Tanya Goyal, Junyi Jessy Li, Greg Durrett. 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] Zhaopeng Feng, Yan Zhang, Hao Li, 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] Shuyin Ouyang, Jie M. Zhang, Mark Harman, 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] Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, 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] Robert Dale. GPT-3: what's it good for?[J/OL]. Nat. Lang. Eng., 2021, 27 (1): 113-118. https://doi.org/10.1017/S1351324920000601.

[13] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, 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] Ying Sheng, Lianmin Zheng, Binhang Yuan, 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] Jianjin Liao, Mingzhen Li, Hailong Yang, 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] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, 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] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, 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] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, 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.

[19] Susan Zhang, Stephen Roller, Naman Goyal, 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] Hugo Touvron, Thibaut Lavril, Gautier Izacard, 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] Khawar Ali. Summary dataset[EB/OL]. 2024. https://huggingface.co/datasets/khwrali011/summary-dataset.

[22] Alessandro Palla. Chatbot instruction prompts[EB/OL]. 2023. https://huggingface.co/datasets/alespalla/chatbot_instruction_prompts.

[23] Gaurang Bharti. Finance alpaca[EB/OL]. 2023. https://huggingface.co/datasets/gbharti/finance-alpaca.

[24] Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, 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.

[25] Github. lightllm[EB/OL]. 2024. https://github.com/ModelTC/lightllm.

[26] Huggingface. Huggingface accelerate[EB/OL]. 2022. https://huggingface.co/docs/accelerate/index.

[27] Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, et al. Distillspec: Improving speculative decoding via knowledge distillation[J/OL]. CoRR, 2023, abs/2310.08461. https://doi.org/10.48550/arXiv.2310.08461. DOI: 10.48550/ARXIV.2310.08461.

[28] Amey Agrawal, Ashish Panwar, Jayashree Mohan, 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.

[29] Felix Brakel, Uraz Odyurt, Ana Lucia Varbanescu. 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.