毕业论文开题答辩记录

## 开题答辩信息

**毕业论文题目**：《大模型KVCache优化方法研究》

**学生姓名**：曾旭阳

**学号**：162140212

**专业**：物联网工程

**指导老师**：蔡淼

**答辩日期**：2025年3月5日14:00 星期三

**答辩地点**：计算机学院203会议室

## 学生陈述

**研究背景与意义**：KVCache技术应用于Transformer模型的Attention机制中，在模型需要计算每个token的Query、Key和Value向量，并通过这些向量来计算Attention权重时，KV Cache技术通过缓存之前步骤计算得到的Key和Value向量，仅对新输入的token进行Attention计算，从而显著降低了冗余计算量。因此，通过对KV Cache的优化，可以显著降低模型的显存占用，提高推理速度，从而使得大语言模型在实际应用中更加高效和经济。

根据KVCache显存占用计算公式（PPT中给出）可知，KVCache的显存占用与模型参数精度字节数、输入/输出序列长度、模型层数、batch size等参数呈线性相关。以GPT-3B模型为例，在长文本（4K）场景下，当batch size成倍增长时，KVCache的显存占用成倍增长甚至达到了模型自身占用的3倍多，严重影响了模型的推理效率。

**研究现状**：目前已有的KVCache的优化方法主要可分四个方向：针对模型的框架优化、计算卸载、采用驱逐策略和存储优化：

Mooncake将预填充阶段和解码阶段解耦，并采用基于预测的早起拒绝策略，在高负载请求场景下有较好的表现；Infinite-LLM采用了一种DistAttention机制，将注意力层与非注意力层分开。这两种方法都适用于GPU集群，对计算资源的需求较高。

PQCahe是在预填充阶段将KV对卸载至CPU，在解码阶段只需要检索需要的KV对即可；PowerInfer则采用一种GPU-CPU混合推理引擎，只计算激活的神经元，从而提高了计算效率。但是当CPU加入计算时，可能会带来新的延迟。

H2O是基于注意力分数动态丢弃低权重的token，从而大幅减少了KVcache的显存占用，但是可能因信息丢失而导致模型性能下降；Ada-KV是采用一种自适应的KVCache驱逐策略，在理论和实验上均证明了其有效性，但是该策略的实时性不足。

PagedAttention将KVCache存储在非连续的显存块中，支持动态分配和释放KVCache，这种方法的显存利用率高，无存储碎片，但是其实现复杂度高。

综上所述，现有的方法的不足可以概括为：缺乏轻量级动态管理、依赖人工规则。

**技术路线和实验设计**：目前预期的技术路线有两种：一种是采用混合精度分层存储，将高频KV头保留为FP16精度，低频头压缩为INT8或FP8，创新点为离线分析注意力头重要性，自适应分配存储精度；一种是KVCache与计算卸载协同优化，将低频KV对从GPU卸载至CPU内存，按需加载，创新点为设计GPU-CPU异步传输流水线，降低数据搬运延迟。

预期实验设计包括长文本生成任务和对话任务，Baseline模型预计采用原生Transformer模型，对比H2O和vLLM(PagedAttention)方法等。

## 教师提问与学生回答

**教师问题1**：计算卸载协同是是谁和谁协同？是有关云边端吗？

**学生回答1**：GPU和CPU协同。因为主要的计算是在GPU上进行，此时CPU相对来说较为空闲，可以将低频的KV对先卸载至CPU上计算，分担GPU的计算压力。

**教师问题2**：协同优化目标是什么？

**学生回答2**：优化目标是降低模型推理延迟，加快模型的推理速度和计算效率。

**教师问题3**：如果只是从对速度的优化角度来说，那么第二个技术路线（计算卸载协同）和第一个技术路线（混合精度分层存储）对精度的优化不是没什么关系吗？（这两者）不需要整合考虑吗？还是说这两条技术路线是只选择一条还是两条都选择？

**学生回答3**：目前想到的优化方向有这两条，具体选择哪一条技术路线现在还未确定。根据我前期的工作，我个人认为第二条技术路线可行性可能高一些。

**教师问题4**：根据你给出的预期实验方案（Baseline模型中包括原生Transformer模型），对于算力的需求很大吗？是否需要训练模型？

**指导老师回答4**：不需要模型训练，只做推理优化任务。使用已经训练好的模型，仅仅对大模型的KVCache机制进行优化。对所以算力要求并不高，单机单卡就可以实现。

## 答辩组意见

**选题评价**：论文题目是《大模型KVCache优化方法研究》，但其实做的是计算卸载协同优化，现题目是否不够具体，优化方法、优化目标可以再具体一点，比如是针对推理精度、推理速度、推理效率等。

**技术路线评价**：开题后，采用的技术路线应尽快确定，不应出现“二选一”的情况。

**答辩组教师**：卢晓珍、卜艳玲、蔡淼

**答辩录音员**：任德翔

**答辩记录员**：曾旭阳