2025 Design, Automation & Test in Europe Conference (DATE 2025)

EVASION: Efficient KV CAche CompreSsion vIa PrOduct QuaNtization

Zongwu Wang[1,2†], Fangxin Liu[1,2†], Peng Xu[1,2], Qingxiao Sun[3], Junping Zhao[4], Li Jiang[1,2,5]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, [2]Shanghai Qi Zhi Institute
[3]SSSLab, Dept. of CST, China University of Petroleum-Beijing, China, [4]Ant Group
[5]·MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{wangzongwu, liufangxin, ljiang_cs}@sjtu.edu.cn

Abstract—**Large language models (LLMs) are increasingly utilized for complex tasks requiring longer context lengths, with some models supporting up to 128K or 1M tokens. This trend, however, presents significant challenges in inference speed and memory management. The primary bottleneck in long-context LLM inference is the quadratic computational complexity of attention mechanisms, causing substantial slowdowns as sequence length increases. KV cache mechanism alleviates this issue by storing pre-computed data, but introduces memory requirements that scale linearly with context length, hindering efficient LLM deployment. Quantization emerges as a promising approach to address the widening gap between LLM size and memory capacity. However, traditional quantization schemes often yield suboptimal compression results for KV caches due to two key factors: i) On-the-fly quantization and de-quantization, causing significant performance overhead; ii) Prevalence of outliers in KV values, challenging low-bitwidth uniform quantization. To this end, we propose EVASION , a novel quantization framework achieving low-bitwidth KV cache through product quantization. First, we conduct a thorough analysis of KV cache distribution, revealing the limitations of existing quantization schemes. Second, we introduce a non-uniform quantization algorithm based on product quantization, which efficiently compresses data while preserving accuracy. Third, we develop a high-performance GPU inference framework for EVASION that leverages sparse computation and asynchronous quantization, significantly enhancing inference speed. Comprehensive evaluation results demonstrate that EVASION can achieve 4 bits quantization trivial perplexity and accuracy loss.**

I. Introduction

Large language models (LLMs) have shown significant capabilities, with recent advancements focusing on extending context lengths to enable applications like long document summarization and extended multi-turn interactions. However, longer contexts introduce substantial computational demands, scaling quadratically with context length [1]–[5]. To mitigate this, modern LLMs use key-value (KV) caches, reducing complexity to $O(n)$ but increasing memory usage. Batch processing further exacerbates storage overhead, as KV caches cannot be shared between requests. For example, a 540B parameter model with a batch size of 512 and context length of 2048 can require 3TB of memory for KV caches alone, creating a memory bottleneck during decoding.

Quantization, successful in weight compression, faces challenges with KV caches due to outliers and dynamic generation. Recent methods like KIVI [6], and KVQuant [7] improve low-bitwidth accuracy but often degrade inference speed.

This work introduces EVASION, a novel quantization scheme for efficient dynamic KV compression. It uses non-uniform quantization based on Product Quantization (PQ) [8] to handle outliers and supports mixed-precision quantization. An efficient GPU inference framework is designed to accelerate processing through sparse computation and asynchronous quantization, reducing memory overhead without significant performance loss.

In summary, we make the following contributions:

- **KV distribution analysis:** Identifies limitations of existing quantization methods and challenges posed by outliers and dynamic KV generation.
- **PQ-based non-uniform quantization:** Proposes a novel scheme using PQ to distribute quantization power unevenly, enabling efficient compression with high accuracy.
- **Efficient GPU inference system:** Designs a high-performance framework for EVASION , accelerating inference and reducing memory overhead through sparse computation and asynchronous quantization.

II. Method and Results

A. Motivation

In this section, we analyze the distribution of KV cache to reveal the limitations of existing low-bitwidth quantization methods. This insight informs our proposed product quantization-based scheme for efficient KV cache compression.

1) Outliers hinder low-bitwidth quantization

To explore the root cause of poor performance in existing KV quantization algorithms, we analyzed the KV distributions across different models, as shown in Fig. 1. Results show that key cache outliers are concentrated in specific channels, while value cache outliers lack anisotropic distribution. Outliers expand the numerical range, complicating low-bitwidth quantization, as accuracy depends on distribution range. Group-wise and mixed-precision quantization can address outliers but introduce additional quantization and de-quantization overheads.

2) Exploring Opportunities for Non-uniform Quantization

† Both authors contributed equally to the paper. This work is supported by the National Natural Science Foundation of China (Grant No.62402311) and Natural Science Foundation of Shanghai (Grant No.24ZR1433700). Li Jiang is the corresponding author.

978-3-9826741-0-0/DATE25/© 2025 EDAA

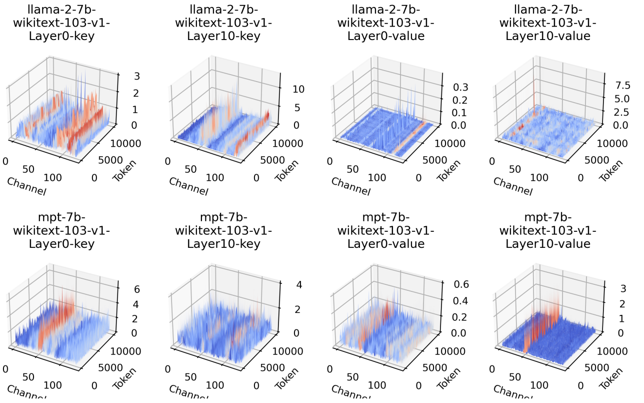

Fig. 1. Magnitude distribution of key and value cache for Llama-2-13B and Falcon-7B.

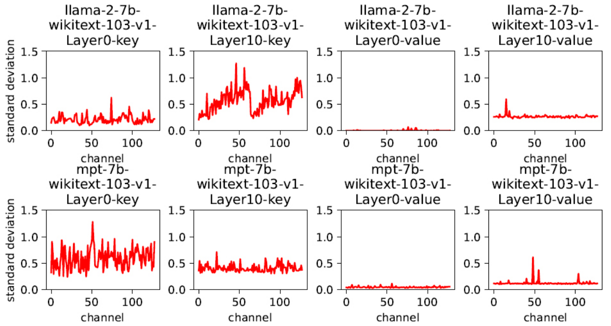

Fig. 2. channel-wise standard deviation distribution of key and value cache for Llama-2-13B and Falcon-7B.

We further explore the standard deviation distribution of KV across channels in Fig. 2, revealing that key standard deviations exhibit outliers in specific channels, while value standard deviations remain small. This indicates that key quantization is more challenging than values, requiring varying precision per channel. Given amplitude and standard deviation outliers in KV cache, we propose using Product Quantization (PQ) for compression.

B. *EVASION Framework Overview*

This section illustrates the algorithm framework of the proposed EVASION, which enhances dynamic KV compression using the PQ algorithm. As shown in Figure 3, EVASION consists of three main components: **offline** PQ codebook training, **online** prefill with KV cache quantization, and **online** decode with KV cache quantization.

Offline PQ Codebook Training. First, KV cache samples are collected, partitioned into sub-vectors, and processed using k-means to generate centroids (codebooks) for each subspace. **Online Prefill with KV Cache Quantization.** In the prefill phase, full-precision keys and values are used for attention computation and then quantized using the trained codebooks. **Online Decode with KV Cache Quantization.** During decoding, attention is computed with all historical tokens by restoring quantized KVs from the cache using the codebooks. The current key and value are quantized and appended to the KV cache for future steps.

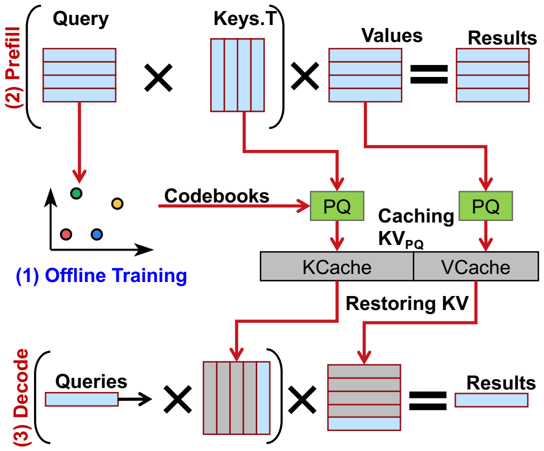

Fig. 3. An overview of EVASION algorithm framework.

TABLE I
PERPLEXITY EVALUATION RESULTS ON VARIOUS MODELS ACROSS DIFFERENT DATASETS.

Method	GPT2-xl		LLaMA-2-7B			MPT-7B	
	Wikitext-2	PTB	Wikitext-2	PTB	C4	Wikitext-2	PTB
baseline	17.41	21	5.12	28.31	7.18	7.68	9.98
KVQuant-3b	18.59	23	11.21	12323.75	16.78	27681432	3E+07
KVQuant-3b-1%	18.24	22.35	5.22	24.34	7.31	7.826	10.14
EVISION-3b	17.6	21.14	5.2	29.55	-	7.79	10.08
KVQuant-4b	18.156	22.32	6.99	102.21	11.02	24043348	3E+07
KVQuant-4b-1%	18.108	22.15	5.14	25.86	7.21	7.71	10.02
EVISION-4b	17.57	21.12	5.23	29.56	7.32035	7.75	10.03

C. *Results*

In this section, we conduct experiments and compare EVASION with existing methods. We experimentally show that EVASION achieves both effectiveness and efficiency(Outlier immunity).

III. CONCLUSION

We present EVASION, a novel and efficient PQ-based quantization framework for LLM inference. EVASION efficiently compresses data while preserving accuracy. It incorporates asynchronous quantization, enhancing inference speed and reducing additional operations caused by hardware-inefficient quantization/dequantization for KV cache compression. Experiments demonstrate that EVASION achieves better model performance while maintaining low latency.

REFERENCES

[1] anthropic, "Introducing claude 3.5 sonnet \ anthropic," https://www.anthropic.com/news/claude-3-5-sonnet, (Accessed on 07/09/2024).

[2] openai, "Models - openai api," https://platform.openai.com/docs/models, (Accessed on 07/09/2024).

[3] Y. Chen *et al.*, "Longlora: Efficient fine-tuning of long-context large language models," *arXiv preprint arXiv:2309.12307*, 2023.

[4] F. Liu *et al.*, "Spark: Scalable and precision-aware acceleration of neural networks via efficient encoding," in *HPCA*. IEEE, 2024, pp. 1029–1042.

[5] F. Liu, N. Yang *et al.*, "Inspire: Accelerating deep neural networks via hardware-friendly index-pair encoding," in *DAC*, 2024, pp. 1–6.

[6] Z. Liu *et al.*, "Kivi: A tuning-free asymmetric 2bit quantization for kv cache," *arXiv preprint arXiv:2402.02750*, 2024.

[7] C. Hooper *et al.*, "Kvquant: Towards 10 million context length llm inference with kv cache quantization," 2024. [Online]. Available: https://arxiv.org/abs/2401.18079

[8] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *PAMI*, vol. 33, no. 1, pp. 117–128, 2010.