ChunkKV: Semantic-Preserving KV Cache Compression for Efficient Long-Context LLM Inference

Xiang Liu [1] Zhenheng Tang [2] Peijie Dong [1] Zeyu Li [1] Bo Li [2] Xuming Hu [1] Xiaowen Chu [1]

Abstract

To reduce memory costs in long-context inference with Large Language Models (LLMs), many recent works focus on compressing the key-value (KV) cache of different tokens. However, we identify that the previous KV cache compression methods measure token importance individually, neglecting the dependency between different tokens in the real-world language characterics. In light of this, we introduce ChunkKV, grouping the tokens in a chunk as a basic compressing unit, and retaining the most informative semantic chunks while discarding the less important ones. Furthermore, observing that ChunkKV exhibits higher similarity in the preserved indices across different layers, we propose layer-wise index reuse to further reduce computational overhead. We evaluated ChunkKV on cutting-edge long-context benchmarks including LongBench and Needle-In-A-HayStack, as well as the GSM8K and JailbreakV in-context learning benchmark. Our experiments with instruction tuning and multi-step reasoning (O1 and R1) LLMs, achieve up to 10% performance improvement under aggressive compression ratios compared to existing methods.

1. Introduction

Large Language Models (LLMs) have become essential for addressing various downstream tasks of natural language processing (NLP), including summarization and question answering, which require the interpretation of a long context from sources such as books, reports, and documents, often encompassing tens of thousands of tokens (Brown et al., 2020; Tay et al., 2022; Touvron et al., 2023). Recent advances in long-context technology within the field of machine learning (ML) systems (Dao, 2024; Jacobs et al., 2023; Xiao et al., 2024) have significantly enhanced computational throughputs and reduced latency of LLMs to process increasingly large input context lengths (Liu et al., 2024b; Young et al., 2024) with saving historical KV cache (key value attentions). However, the memory requirement of the KV cache in serving super-long contexts becomes a new bottlneck (Zhang et al., 2023; Reid et al., 2024). For instance, the KV cache for a single token in a 7B-parameter model requires approximately 0.5 MB of GPU memory, resulting in a 10,000-token prompt consuming around 5 GB of GPU memory.

To address the substantial GPU memory consumption caused by KV caching, recent studies consider compressing the KV cache by pruning non-important discrete parts from the prompt tokens (Zhang et al., 2023; Li et al., 2024; Ge et al., 2023; Cai et al., 2024; Fu et al., 2024a; Yang et al., 2024b; Liu et al., 2024e; Tang et al., 2024). H2O (Zhang et al., 2023) and SnapKV (Li et al., 2024) have shown that retaining less than 50% of the discrete KV cache can significantly reduce GPU memory usage with minimal impact on performance. However, we identify that the previous KV cache compression methods (Zhang et al., 2023; Cai et al., 2024) measure token importance isolatedly, neglecting the dependency between different tokens in the real-world language characterics. For example, as shown in Figure 1, focusing on token-level importance might excessively focus on words about subjects "turaco" in the question while omitting crucial information about the objects (foods) in the documents, resulting the loss of essential semantic information. This motivates us to rethink the following question:

How to avoid isolated token importance measurement and preserve the semantic information in KV cache?

In light of this, we observe that the complete semantic information usually appear in a continuous sequence (Fang & Xie, 2022). Thus, we introduce a straightforward yet effective ChunkKV, grouping the tokens in a chunk as a basic compressing unit, which should be preserved or discarded as a whole. Thus, it retains the most informative **semantic chunks** from the original KV cache. As shown in Figure 1, preserving a chunk helps to catch the subject, predicate, and object. Furthermore, we investigate that *the preserved*

[1] The Hong Kong University of Science and Technology(Guangzhou), Guangzhou, China [2] The Hong Kong University of Science and Technology, Hong Kong, China. Correspondence to: Xuming Hu <xuminghu@hkust-gz.edu.cn>, Xiaowen Chu <xwchu@hkust-gz.edu.cn>.