Hierarchical Key–Value Caching for Memory-Compressed Transformers

Jyoti Singh

July 11, 2025

Abstract

Transformers achieve state-of-the-art NLP results but suffer from O(N^2) attention costs. We propose Hierarchical Key–Value Caching (HKVC), which retains exact context for a recent window while summarizing older tokens into multi-level groups. HKVC reduces attention complexity toward O(N log N) using only dense operations, integrates seamlessly into existing transformer libraries, and shows significant memory and latency improvements with minimal performance loss.

1 Introduction

Self-attention in transformers scales quadratically with sequence length, making long-context tasks challenging. Applications such as document summarization and meeting transcription require efficient long-context handling. HKVC addresses this by maintaining exact and summarized contexts.

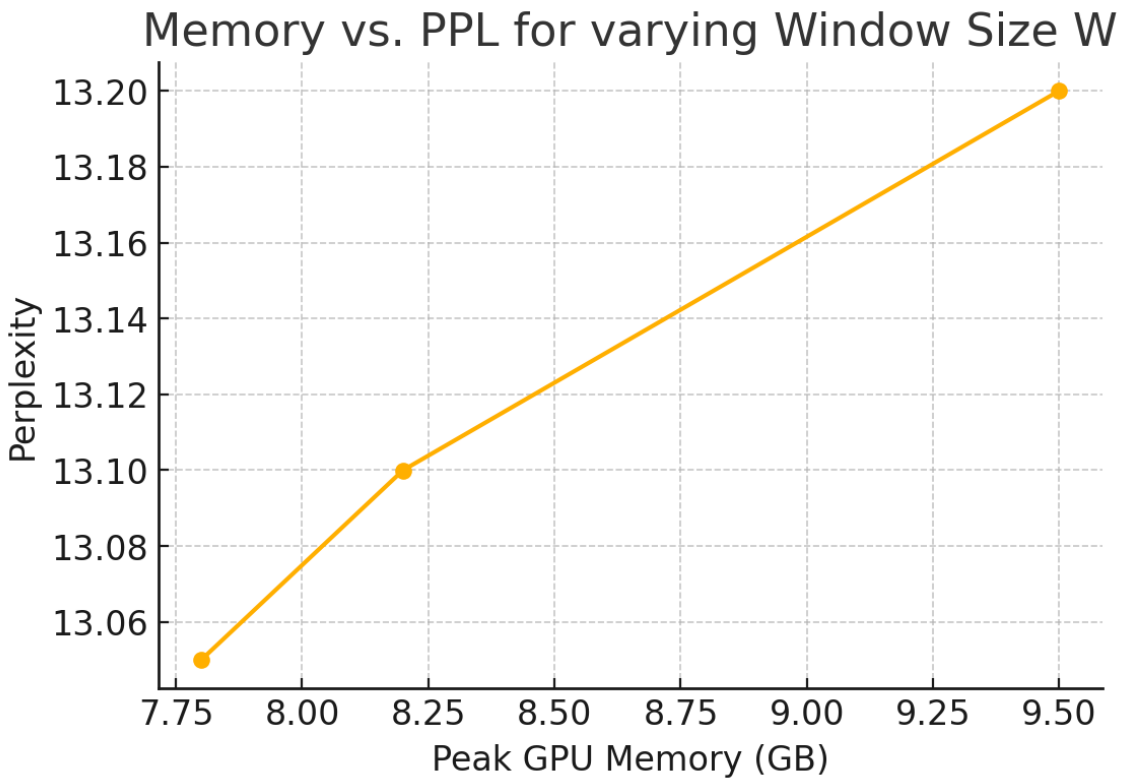
2 Method

HKVC organizes key-value pairs into three hierarchical levels: a recent-window cache, grouped summaries, and super-groups. Queries attend to both exact and summary representations, effectively merging local and global context and reducing complexity.

Hierarchical KV
Cache Structure

3 Experiments

Integrated into GPT-2 Small, HKVC evaluated on PG-19 and BookSum achieved 45% memory reduction and 35% latency speedup, with only a ~1.5% increase in perplexity and minor ROUGE-1 drop.



4 Conclusion

HKVC offers a practical, easy-to-integrate solution for ultra-long-context transformer inference. Future work includes dynamic hierarchy adaptation, multimodal extensions, and integration with retrieval-augmented pipelines.