Qeios

Peer Review

Review of: "SimLayerKV: A Simple Framework for Layer-Level KV Cache Reduction"

Yao Chen[1]

1. Computer Science, National University of Singapore, Singapore

Summary:

The paper introduces SimLayerKV, a method for reducing Key-Value (KV) cache redundancy in large language models (LLMs) by selectively trimming the cache in identified "lazy" layers. The approach is training-free, easily implemented, and demonstrates promising results on several LLMs across a variety of tasks. The method identifies lazy layers based on attention weight patterns and reduces their KV cache accordingly. The experiments show that SimLayerKV achieves a 5x KV cache compression ratio with a small performance drop, especially when combined with 4-bit quantization.

Strong Points:

1. The proposed solution is attractive, and the solution seems effective based on the given evaluations.

2. The method is training-free and integrates smoothly into popular inference frameworks, making it practical and easily adoptable.

Weak Points:

1. Limited Model Size Evaluation: The paper primarily focuses on models with similar sizes (7B and 8B parameters), and the paper could be further enhanced by exploring a wider range of model sizes, from smaller models to much larger models. The effectiveness of the proposed solution on different model sizes should also be evaluated.

2. Insufficient Quantization Evaluation: The paper primarily evaluates SimLayerKV in combination with 4-bit quantization. While 4-bit quantization is useful for reducing memory usage, the paper could explore other quantization levels (e.g., 8-bit, 6-bit, or even lower bits). It is unclear how the method will perform with different quantization levels. The effect on performance with different quantization levels should be investigated.

 doi.org/10.32388/8XXTP7

Declarations

Potential competing interests: No potential competing interests to declare.