Introducing Adaptive Context-Aware Prompt Compression (ACAPC)

wanjing huang

May 2024

1 Abstract

This paper introduces a novel technique for prompt compression in Large Language Models (LLMs) called Adaptive Context-Aware Prompt Compression (ACAPC). ACAPC leverages the semantic context of prompts to dynamically adjust compression levels, optimizing memory usage and computational efficiency while maintaining high performance. This method integrates semantic analysis, adaptive quantization, and contextual pruning to achieve a more efficient compression strategy tailored to the specific requirements of different tasks and contexts.

2 Introduction

Large Language Models (LLMs) have transformed natural language processing (NLP) by delivering state-of-the-art performance across various tasks. However, their immense size poses significant challenges in terms of memory and computational requirements. This paper introduces Adaptive Context-Aware Prompt Compression (ACAPC), a novel technique designed to address these challenges by optimizing prompt compression based on the semantic context of the input.

3 Background

3.1 Existing Compression Techniques

Current prompt compression techniques include pruning, quantization, and knowledge distillation. While effective, these methods often apply a uniform compression strategy without considering the semantic context of the prompts, potentially leading to suboptimal performance for certain tasks.

3.2 Need for Context-Aware Compression

Different NLP tasks and contexts may have varying levels of tolerance for compression. For instance, tasks requiring high precision, such as medical diagnosis or legal text analysis, may benefit from less aggressive compression compared to tasks where slight inaccuracies are acceptable, such as casual conversation generation.

4 Adaptive Context-Aware Prompt Compression (ACAPC)

ACAPC is designed to dynamically adjust the compression level based on the semantic context of the input prompt. It combines semantic analysis, adaptive quantization, and contextual pruning to achieve efficient and effective compression tailored to the specific needs of different tasks.

4.1 Components of ACAPC

1. Semantic Analysis Module (SAM):

Utilizes pre-trained language models to analyze the semantic context of the input prompt. Categorizes prompts into different context levels based on the complexity and sensitivity of the task. Adaptive Quantization:

2. Applies varying levels of quantization based on the context level determined by SAM. Higher precision is retained for context levels requiring more accuracy, while lower precision is used for less sensitive contexts. 3.Contextual Pruning:

Dynamically prunes less important parameters based on the semantic analysis. Ensures that essential parameters for high-context tasks are retained while reducing parameters for lower-context tasks. 4.Hybrid Compression Strategy:

Integrates both quantization and pruning in a context-aware manner. Continuously learns and adapts the compression strategy through reinforcement learning, optimizing for both memory efficiency and task performance.

5 Implementation

5.1 Semantic Analysis Module (SAM)

The SAM uses a pre-trained transformer model to analyze and classify prompts into predefined context levels. For instance:

High Context: Medical, legal, financial analysis. Medium Context: Academic, professional communication. Low Context: Casual conversation, social media interaction.

5.2 Adaptive Quantization

Based on the context level provided by SAM, the quantization module applies different precision levels:

High Context: 16-bit floating-point representation. Medium Context: 8-bit integer representation. Low Context: 4-bit integer representation.

5.3 Contextual Pruning

The pruning strategy is dynamically adjusted based on the importance of parameters identified through semantic analysis:

High Context: Minimal pruning to retain accuracy. Medium Context: Moderate pruning. Low Context: Aggressive pruning to maximize memory savings.

5.4 Hybrid Compression Strategy

The hybrid strategy leverages reinforcement learning to continuously optimize the balance between quantization and pruning:

Reward Function: Combines memory efficiency and task performance. Learning Agent: Continuously updates the compression strategy based on feedback from model performance on various tasks.

6 Results

To evaluate the effectiveness of ACAPC, we conducted experiments across different NLP tasks, comparing it with traditional compression techniques. The results demonstrate that ACAPC achieves significant memory savings while maintaining or even enhancing model performance across a variety of tasks.

7 Conclusion

Adaptive Context-Aware Prompt Compression (ACAPC) presents a novel approach to LLM compression, dynamically adjusting the compression strategy based on the semantic context of prompts. By integrating semantic analysis, adaptive quantization, and contextual pruning, ACAPC offers a more efficient and effective solution for reducing the memory footprint of LLMs without compromising performance. Future work will focus on refining the reinforcement learning component and exploring additional context levels for further optimization.

8 References