論文内容の要旨

申請者氏名　　GUO ZHIYU

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), powering applications such as conversational agents, programming assistance, and long-document summarization. However, their practical deployment faces significant challenges: the high inference costs in long-context scenarios and the substantial memory demands of Multi-Layer Perceptron (MLP) modules, which dominate model parameters. This dissertation tackles these issues with two novel methods to enhance inference efficiency while maintaining performance.

First, Value-Aware Token Pruning (VATP) is proposed for KV cache reduction. By incorporating both attention scores and the L1 norm of value vectors to evaluate token importance, VATP addresses the limitations of conventional approaches that rely solely on attention scores. Extensive experiments on LLaMA2-7B-chat and Vicuna-v1.5-7B across 16 LongBench tasks demonstrate that VATP outperforms attention-score-only baselines in over 12 tasks.

Second, Dependency-Aware Semi-structured Sparsity (DaSS) is introduced to compress GLU-based MLP modules. DaSS incorporates structural dependency into the weight magnitude-based unstructured pruning. Evaluations on LLaMA2, Mistral, and Gemma models show that DaSS surpasses state-of-the-art methods SparseGPT and Wanda in hardware-friendly N:M sparsity while maintaining computational efficiency.

Collectively, these contributions significantly enhance LLM inference efficiency, paving the way for broader adoption and deployment of these powerful models in diverse real-world applications.

論文審査結果の要旨

申請者氏名　　GUO ZHIYU

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), powering applications such as conversational agents, programming assistance, and long-document summarization. However, their practical deployment faces significant challenges: the high inference costs in long-context scenarios and the substantial memory demands of Multi-Layer Perceptron (MLP) modules, which dominate model parameters. This dissertation tackles these issues with two novel methods to enhance inference efficiency while maintaining performance.

First, Value-Aware Token Pruning (VATP) is proposed for KV cache reduction. By incorporating both attention scores and the L1 norm of value vectors to evaluate token importance, VATP addresses the limitations of conventional approaches that rely solely on attention scores. Second, Dependency-Aware Semi-structured Sparsity (DaSS) is introduced to compress GLU-based MLP modules. DaSS incorporates structural dependency into the weight magnitude-based unstructured pruning. These contributions significantly enhance LLM inference efficiency, paving the way for broader adoption and deployment of these powerful models in diverse real-world applications.

The two studies are published as one high quality peer-reviewed journal paper and one peer-reviewed international conference paper. The research has an engineering contribution in efficient inference for LLMs and would have an impact not only to natural language processing, but to the relevant fields of machine learning, e.g., text generation. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.