Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Denis Kuznedelev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Accurate and efficient methods for neural network compression

Supervisor: Professor Dmitry Yarotsky

Name of the Reviewer: Ivan Oseledets

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 01-12-2024**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the

Reviewer's Report

The thesis by D. Kuznedelev is very well-written and contains very motivating results published at top conferences. Below is the summary of the main results, comments on the novelty and some critical comments.

Summary of the main results:

The thesis presents several novel algorithms and techniques for compressing deep neural networks, focusing on two approaches: pruning (sparsification) and quantization. The key contributions include:

1. CAP: A correlation-aware pruning framework that achieves state-of-the-art results on compressing modern vision models like ViTs and ConvNets, outperforming previous methods by a significant margin.

2. Analysis of the difficulty of sparse optimization and training, both for vision and language models. The author provides evidence that standard training recipes are insufficient, and proposes specialized techniques to address this, setting new benchmarks.

3. Sparse fine-tuning methods for accelerating large language models, leveraging knowledge distillation to recover accuracy at high sparsity levels.

4. Theoretical analysis of the Iterative Optimal Brain Surgeon (I-OBS) algorithm, providing convergence guarantees and connections to prior work.

5. Novel quantization approaches for LLMs, including the SpQR method that isolates and compresses outlier weights separately, and the AQLM algorithm that adapts additive quantization to language models.

Novelty and Significance:

The thesis makes several novel contributions to the field of model compression. The CAP pruning framework, the sparse optimization analysis, and the I-OBS algorithm with theoretical guarantees are important advances that push the state-of-the-art. The LLM compression techniques, especially AQLM, demonstrate new ways of adapting quantization methods to the unique challenges of large language models. Overall, the work provides a comprehensive study of compression techniques, with a good balance between theoretical analysis and practical, high-performing algorithms.

Thesis Structure:

The thesis is well-structured, starting with an introduction and background on fundamental concepts. It then dedicates individual chapters to the key contributions, each providing detailed explanations, experiments, and discussions. The overall flow logically builds up from pruning to quantization, gradually increasing the complexity of the addressed problems. The appendices provide additional technical details and proofs, which complement the main text well.

Comments:

1) Two techniques are considered, sparsity and pruning, in many variants. Given a large model (LLM) what is the optimal way of reducing its size?
2) Can sparsity and quantization be combined?
3) A lot of models now are already trained in low precision (8 bit, or even 1bit-LLM). Do you believe that quantization methods will be used for those cases?

Overall, I think this is an excellent work.

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense

I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report

The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense