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:

I confirm the absence of any conflict of interest	**Date: 16-11-2024**

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 Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications
- The summary of issues to be addressed before/during the thesis defense

This dissertation is a thorough and detailed study about neural network compression. It involves exploration of existing techniques for neural network compression - pruning, quantization, and knowledge distillation—identifies their limitations and proposes novel theoretically motivated compression algorithms with provable performance in various applications. The thesis includes 7 publications, 5 of which were accepted to Core A* conferences (NeurIPS, ICLR, ICML). Therefore, I believe that this thesis is a strong work with a pronounced scientific contribution.

The papers involved in the thesis incorporate various ideas to advance further the compression/quality frontier. Specifically, CAP, I-OBS pruning algorithms, and SpQR, AQLM quantization algorithms leverage second order Taylor decomposition to estimate the importance of parameters and find optimal compressed weight configuration. AQLM adopts a vector quantization approach for an accurate approximation of original weight matrices in LLM (Large Language Models). The aforementioned approaches show state-of-the-art results when applied to compression of CNN/ViT architectures adopted in computer vision and LLM.

The aforementioned conferences (NeurIPS, ICLR, ICLM) are leading conferences on machine learning and artificial intelligence (Core A*). The peer-review process is strict there, and only submissions with enough novelty, pronounced scientific contribution are accepted. Hence, the publications included in the thesis comply with international standards and provide a nontrivial contribution to the domain.

Results obtained have an apparent practical application, as the main goal of the study is the reduction of deployment and maintenance costs of advanced AI technology in a plethora of applications. Specifically, the introduced algorithms can be applied for vision systems in autonomous devices and inference of LLMs on consumer-grade PCs and mobile devices. Inference on local hardware is often highly desirable due to personal data privacy concerns, and the presented results provide a potential for widespread adoption and deployment on the edge.

This thesis shows diligent and extensive Ph.D. student work. However, sometimes the manuscript turns out to be hard to read due to the presence of a large amount of information about experimental setup and a bit of heterogeneous structure. Nevertheless, it is likely that incorporation of the provided results, comments, figures and tables is necessary for a complete presentation. While reading the text, I've identified minor grammar, syntax errors and typos, and I would recommend the student double-check the text once again and fix them. Overall, these do not hurt the readability of the thesis text.

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*