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: Alexey Frolov

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**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

The thesis is overall well-written with clear motivation, structure, and logical flow. The background and related work are sufficient for one, not an expert domain, to grasp the basics and learn about the current state-of-the-art in the field of neural network compression. Each paper involved in the dissertation targets a specific problem, analyzes limitations and points missed by previous works, and presents a highly performant solution. The key points and concepts are present in the main text, whereas technical details and theorem proofs are deferred to appendices. In my opinion, the results obtained in the thesis are quite motivating and valuable to the community.

This thesis provides a study of several dominant techniques applied to neural network compression - pruning, quantization, and knowledge distillation. The algorithms introduced in the work have a solid theoretical footing - namely, Taylor expansion of the loss function to estimate the importance of each parameter and find an optimal compressed configuration. Additive vector quantization, introduced in AQLM paper, has a strong motivation for low-bit compression, as it is known that vector quantization being more flexible can provide a tighter approximation of the target data distribution. All papers included in the thesis manuscript target the problem of model compression leveraging various ideas from analysis, linear algebra, and optimization, therefore, I believe that the topic of the dissertation is highly relevant to the content

The thesis involves 7 papers, 5 of them were published at the leading machine learning conferences (Core A*):

- CAP - NeurIPS 23
- SpQR - ICLR 2024
- AQLM - ICML 2024
- PV-Tuning - NeurIPS 2024
- I-OBS NeurIPS 2024

These conferences have a very high impact and are known for the strict review process and high competition. Hence, I am confident that these publications are of sufficient novelty and comply with international standards. One of the papers (SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression) was cited more than 100 times.

The results obtained in this thesis have an immediate practical application. Pruning and fine-tuning algorithms designed for computer vision models presented in this work may significantly reduce the operation cost and energy consumption of neural networks deployed on edge and mobile devices. Quantization algorithms for Large Language Models pave the way for inference of modern language models (Llama, Qwen, Gemma, Mistral) on consumer-grade PCs and laptops.

All in all, I think that student did a good job during his Ph.D. study, and the results obtained provide a pronounced contribution to the domain and noticeably advance current state-of-the-art. The thesis is a bit "bulky", making it sometimes challenging to read. Yet the manuscript has a clear logical flow; therefore, it is likely that one cannot significantly compress the thesis without compromising the completeness of the discussion.

There are some minor grammar and syntax issues in the text of the manuscript - missing articles, forgotten commas, hyphens.

A question one would be interested to have an answer to - what are the actual limits of the neural compression with respect to pruning and quantization, and their compound application? And how does compressibility depend on the task and the model?

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*