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 Naumov

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 02-12-2024**

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 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

Brief evaluation of the thesis quality and overall structure of the dissertation

This thesis is a solid work with significant scientific contribution. The manuscript is well-structured, starting with an exhaustive introduction and background on the topic. The importance and significance of the problem to the research community and practitioners is clearly stated. Next, the thesis provides a detailed exploration of different techniques to improve the efficiency and accuracy of existing model compression techniques and presents a detailed description of the presented approaches. Lastly, the thesis provides a summary of the obtained results and directions for further study.

The relevance of the topic of dissertation work to its actual content

The topic of dissertation is highly relevant to the content. This work is stated as a study devoted to improving and advancing current state-of-the-art in the domain of neural network compression and the works involved in the thesis explore this problem from different angles, leveraging sparsity, quantization, knowledge distillation as dominant approaches in the field.

The relevance of the methods used in the dissertation

In this dissertation various compression paradigms - sparsity, quantization, accompanied with knowledge distillation were adopted. These are the most common and successful methods for neural network compression; therefore, I believe that these are appropriate and relevant to the work. In addition, most of the papers involved in this study leveraged the Optimal Brain Surgeon (OBS) framework, based on second order Taylor decomposition, to estimate the importance of parameters and determine optimal compressed weight configuration. It is a theoretically motivated approach with proven strong empirical performance.

The scientific significance of the results obtained and their compliance with the international level and current state of the art

This thesis is based on seven papers, five of which were published at top machine learning conferences, specifically, NeurIPS 2023, ICLR 2024, ICML 2024, NeurIPS 2024 (all Core A*). NeurIPS, ICLR, ICML are the three primary conferences with the highest impact on machine learning and artificial intelligence research. Therefore, I believe that the results are of scientific significance, comply with international standards and align with current state-of-the-art.

The relevance of the obtained results to applications (if applicable)

The obtained results are highly practical, targeting the reduction of inference and deployment costs of advanced AI technologies for both practitioners and common users. Specifically, methods developed in this work can be used for inference of computer vision systems on edge devices and local inference of modern large language models (LLMs).

The quality of publications

The publications included in the thesis were presented at leading conferences on machine learning and artificial intelligence (Core A*). These venues are known for their strict peer-review processes and high competition, ensuring that only research with sufficient novelty and scientific impact is accepted. All papers

involved have a clear problem statement and scientific challenge and provide a method to solve the raised problem. Each method introduced comes with an exhaustive experimental assessment and detailed ablation. Overall, the publications satisfy high international standards and provide significant contributions to the domain.

The summary of issues to be addressed before/during the thesis defense

Overall, this thesis is a solid work.

The main issue of this work is the abundance of technical and experimental details that may be presented in the appendices to improve readability and information flow. This work contains a lot of different results; therefore, I highly recommend putting them under the same "roof" to make the presentation more accessible to the audience on defense. In addition, some parts of the work may be hard to grasp without an illustration, such as the concept of vector quantization, alternating optimization of continuous/discrete parameters. Adding some schematic visualization could be beneficial to 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*