Skoltech

Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Viktoriia Chekalina

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Computationally efficient Natural Language Processing methods using tensor representations

Supervisor: Associate Professor Alexander Panchenko

Name of the Reviewer: Alexey Frolov

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	[redacted] **Date: 20-08-2023**

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 PhD thesis of Viktoriia Chekalina, entitled “Computationally efficient Natural Language Processing methods using tensor representations”, is devoted to important and challenging problems related to the Natural Language Processing (NLP) area. Indeed, the increased hardware capabilities and the invention of new architectures such as transformer models led to significant progress in NLP area. Current NLP models show an excellent performance, but the clear drawbacks are as follows: (a) the model is large, i.e., it should have huge number of trainable parameters to demonstrate good performance; (b) the model should be trained on big corpus of texts. This thesis addresses both problems. Viktoriia proposes to utilize low-rank representations to compress either the architectures or the dataset itself. Such an approach allows to reduce the number of trainable parameters without performance degradation. The presented results are new and original and supported by publications.

The thesis consists of an introduction, 6 chapters and a conclusion. In the Introduction Viktoriia explains the motivation, states the research problem and main objectives, and justifies novelty, contribution, and impact. In my opinion, it would be worthy to briefly consider state-of-the-art methods for NLP models compression and mention their drawbacks.

Chapter 2 is devoted to the background. In this Chapter Viktoriia describes all the required tools, such as matrix decompositions, tensor decompositions, knowledge graphs and transformer-based language models.

Chapters 3 and 4 are devoted to the developed methods. In Chapter 3 Viktoriia proposes memory efficient knowledge embedding representation model (MEKER), which utilizes canonical polyadic decomposition. This model allows for obtaining optimization gradients without a backpropagation mechanism and reduces the memory needed in training. In Chapter 4 a custom TTM-layer is developed. This layer has smaller number of parameters in comparison to fully connected layer and uses less memory during forward and backward passes.

Chapters 5-7 are devoted to the application of the developed methods to the real-world NLP problems. In Chapter 5 a transformer-based GPT-2 architecture with TTM layers is considered. It is demonstrated that this modification results in a 40% reduction in model size, while maintaining the performance (it is important to mention that such results were achieved for training from scratch). Chapter 6 is devoted to efficient question answering using TTM decomposition. In this Chapter Viktoriia proposed to compress FC layers with use of TTM and SVD decompositions. It appeared that both methods demonstrate similar compression results and highly dependent on the choice of the "proper" layer for compression. The algorithm to choose "proper" layer is also developed. Finally, Chapter 7 is devoted to transformer-based encoders compression using TTM decomposition. Viktoriia tested the performance of compressed BERT and BART models on natural language understanding and generation tasks. It is also important to note that Viktoriia adapted the method proposed by Hsu et al to TTM decomposition by incorporating Fisher information.

Minor critical comments: there is an overlap between Introductions in the Chapters. Do you mean 1100 MB or Mb in Table 4.2?

In my opinion the thesis is well written, with a very clear structure that allows relatively easily following the author’s ideas and developments. Viktoriia obtained significant scientific results compliant with the international level and current state of the art. All results are supported by quite impressive scientific papers: 8 papers in total, 1 paper in A* conference and 2 papers in A conferences.

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*