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:

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 13-08-2023**

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

I liked the overall quality and structure of the thesis. Its topic (utilizing limited resources more efficiently in training machine learning models for NLP task) is highly relevant nowadays. The results achieved by the defendant are important and will undoubtedly be useful for many NLP researchers.

Obviously, the defendant was publishing actively while preparing the thesis, which is a good sign. The list of papers where the defendant is a first author includes ACL'22 Student Research Workshop and EACL'21 System Demonstrations proceedings (among other publications), which is definitely good enough for me.

The thesis is worthy of a public defense. However, I still have to report three important issues and a number of minor ones. These issues (described below) should be taken into account and if possible addressed by the defendant - at least in the presentation during the defense, if time limits do not allow to modify the text itself.

Issues

1. Naming

The thesis title mentions "**Computationally efficient NLP methods using tensor representations**". Was it actually "using tensor compression/decomposition" that was meant?

In section 1.2, the defendant says that one of the thesis research questions is *"Can we reduce the size of the Transformer-based language model by replacing some layers with Tensor structures and how it affects the model performance?"*

But layers of a Transformer (or any artificial neural network, for this matter) are already tensors by definition (of whatever shape). Thus, it feels a bit strange when the defendant is comparing their "tensor-based layers" against "others" as if they are not "tensor-based". This should be explained and made more clear in the Introduction and during the defense.

2. Thesis form

I am not sure what is the intended form of the thesis. Is it a collection of papers or a monograph? It is never stated in the text, but it's obvious that the chapters roughly correspond to the papers authored by the defendant. Among other things, it can be seen by regular slips like "in this paper, we..." and different styles of citations in every chapter - as if the text was copy-pasted without making any effort to adapt LaTeX styles of different conferences to the thesis style file.

Of course there is nothing wrong with a PhD thesis being a collection of papers plus Introduction and Conclusion (given that Skoltech is OK with it). But I would appreciate if the nature of the work was clearly stated in the very beginning. If it is paper-based, it would be helpful to start every chapter with a clear pointer to the paper it is formed of.

If, on the other hand, the thesis is a monograph (which I believe it is not), the author should try to make it a coherent narrative, not a diverse collection of loosely related articles. For example, Chapter 3 feels a bit as an outlier and lacks the explanation of its connections (or necessity) for the subsequent chapters. Section 6.2 would also benefit from a more clear positioning within the wider frame of the thesis.

3. Language

The "Acknowledgments" section expresses gratitude towards those who proofread the thesis, but if fact it abounds in grammar errors and typos which are easy to capture using even the simplest English spellchecker. Unfortunately, this problem is present even in Chapter 1 ("Introduction"), which I would expect to be subject to the strongest scrutiny in terms of text quality.

I definitely recommend the defendant to actually run a spellchecker on the full text of the thesis, and probably hire a professional proofreader. Of course this does not have anything to do with the scientific value of the thesis, but all these typos, repetitions and inconsistencies leave an impression of a hastily compiled text not respecting the reader. In fact, sometimes it reaches the degree when it becomes difficult to even understand what the author wanted to say in a particular paragraph. If need be, I can provide a copy of the thesis PDF file with my comments on the margins.

4. Other minor issues

1. Section 2.4 introduces the notion of Transformer-based language models. Strangely, it does not even give a definition of a language model. The first paragraph just says " *language models possess knowledge about the properties inherent to a language. By taking into account the surrounding text, they are capable of generating a probability distribution of the appearence of each language unit*". This is very vague and I'd suggest to rephrase it making sure of mentioning such notions as "context", "token sequence", "perplexity" and "masked language modeling objective" (instead of "*filling in gaps in the text*").

2. A consistent issue in the text is not mentioning the language of the datasets the author is working with. The language is not specified even for the corpora on which the author trains the GPT-2 model in section 5.4 (only the corpora names and citations, and even this only in the evaluation part, as if language is not important at all). I strongly believe this is not the correct way of introducing data for an **NLP** thesis. We should remember the Bender rule: "Natural Language is **not** a synonym for *English*. Always name the language you are working on". This is important both ethically and practically: statements which are true for English NLP models might not be true when applied to other languages.

3. By the way, as far as I understand, the Wikitext-103 corpus contains only about

100M word tokens in size (it's never stated in the thesis, I had to look it up myself). I am not sure such a small training corpus makes for a good testing ground for a model like GPT-2. At least this issue should be mentioned and at least briefly discussed. The size of the second corpus (OpenWebText) should also be reported in more details than "sufficiently large". Finally, it is not clear why these particular corpora were chosen for evaluating custom TTM layers in GPT-2. Will the results change if trained and tested on other corpora? On other languages?

4. Some equations are numbered and some are not (even within one chapter). It does not look good and does not help with navigating the text.
5. Chapters 5 and 6 result in trained language models. Are these models published anywhere? I would expect them to be available on HuggingFace Model Hub, for example.
6. In the user study in Section 6.2, more data about the demography of human graders should be reported. Who are these people? What is their gender, native language, etc?
7. Table 6.10 should explicitly specify which sub-table represents scores for relevance/quality.
8. I am a bit worried about the examples of retrieval shown in table 6.18. The documents retrieved by the *ColBERT Compressed TTM* model seem to be completely nonsensical. At the same time, the NDCG@5 performance of this model in tables 6.16 and 6.17 is on par with the rest. Is it just a bad choice of an example or...?
9. Tables 7.5 and 7.6 would be much better with labels showing which rows correspond to what type of experiments (task-oriented fine-tuning, compression, further fine-tuning). As of now, the reader only sees three bulks of rows with identical headers on the left, but different values in the cells. This is confusing.
10. Tables 7.7 and 7.8 seem to be in the wrong (swapped) order. At least, the text describes the findings in the inverse order (first summarization, then detoxification).

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*