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: Dr. Artem Shelmanov

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 22-08-2023**

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

Overview

The thesis investigates approaches to reducing the size of graph embeddings and neural network layers based on various tensor decomposition techniques. The main goal is to obtain smaller models with the similar performance that could be trained or run on the hardware with strictly limited capabilities. Chapter 1 discusses the problem and its motivation, as well as the thesis objectives, contributions, and novelty. Chapter 2 gives an overview of matrix decomposition methods and introduces the reader to knowledge graphs (KGs) and Transformer-based language models. Chapter 3 presents MEKER – a method for knowledge graph embedding based on the canonical polyadic decomposition and its evaluation results. Chapter 4 discusses an approach to reduction of fully connected (FC) layers in a Transformer model based on tensor train matrix (TTM) containers, as well as efficient implementations of the forward and backward passes. It also discusses the time and memory complexity of various decompositions and provides empirical evaluations of their resource-efficiency. In Chapter 5, the author adapts TTM-based FC layers to GPT models and provides experimental evaluation of the modified model on language modelling and language understanding tasks. Chapter 6 provides a discussion and evaluation of a comparative question-answering (QA) system, which seems somewhat unaligned with the main topic of the thesis. However, this chapter also presents some results for a ranking Transformer modified with TTM-based FC layers. Chapter 7 introduces a modification of the TTM decomposition for FC layers and includes experiments on the natural language understanding (NLU) task, wherein the author compares various tensor decomposition methods applied to the FC layers of a Transformer-based encoder. Chapter 8 summarizes and concludes the work.

Motivation of the problem and relevance of the suggested methods

The problem raised in the thesis is crucial for the development of modern large language models, as hardware is a common limitation for deploying neural models in real-world applications. It also represents a serious bottleneck for researchers. The author has chosen to tackle this issue by investigating tensor decomposition methods, which I find a fascinating research direction filled with many unsolved problems and practical obstacles.

Obtained results compared to state of the art

The obtained results are a valuable contribution to the research field when considering the tensor decomposition for neural networks per se. There are four major results:

1. MEKER – a model for KG embedding that aims to reduce its memory consumption and enables training on GPUs with a small RAM limit. MEKER suggests using canonical polyadic decomposition for representing the embedding matrix and an analytical implementation of the gradient calculation procedure for computationally efficient training.

   According to experiments, MEKER is somewhat in the middle between two methods from previous work: QuatE (Zhang et al., 2019) and ComplEX (Trouillon et al. 2016). QuatE achieves better results in the link prediction task but at the expense of higher memory consumption during training, due to its 4-dimensional representations. In contrast, ComplEX obtains slightly better memory consumption while usually gives slightly worse or similar results compared to MEKER. MEKER slightly outperforms ComplEX on the KG QA task on the RuBQ benchmark. Other considered methods are substantially inferior to MEKER in terms of performance and/or memory consumption.

2. Method for the compression of fully connected layers using tensor train matrix (TTM) containers.
    a. An approach for decomposing weight matrices in FC layers based on TTM containers.
    b. A computationally efficient algorithm for the forward pass that alleviates the overhead due to increased number of multiplications during inference of a TTM-decomposed layer (based on the optimized tensor contraction schedule).
    c. A computationally efficient algorithm for training of such layers that performs backward pass with an optimized tensor contraction schedule.

   The thesis includes an analysis of time and memory complexity, as well as empirical justification of the proposed solution.

3. A modification of a TTM-based FC layer that incorporates Fisher information into decomposition algorithms, featuring a two-step model fine-tuning approach (FWTTM). Fisher information is used to reweight the importance of the parameters in the FC layer matrix reconstruction objective. An additional fine-tuning step is introduced to improve model performance after the weights in FC layers are decomposed. The proposed modification results in performance improvements compared to the original TTM-based layer, particularly in NLU and text generation tasks.

4. Recommendations for using TTM-based FC layers in Transformer models (GPT-like and BERT-like):
    a. Recommendations for selecting hyperparameters of TTM containers suitable for Transformer models.
    b. Recommendations regarding which FC layers in Transformers could be replaced with TTM-based layers (for ColBERT and BERT).

   The provided recommendations allow to achieve descent trade-offs between model compression and performance drop in language modelling, natural language understanding, and text generation tasks compared to other approaches to the matrix decomposition like SVD. However, it is worth noting that the results usually do not demonstrate benefits compared to other compression techniques based on distillation and reduced precision (fp16) weight representation.

The relevance of the obtained results to applications

The thesis does not provide sufficient evidence to suggest that the proposed methods based on tensor decomposition are practical for use in real-world applications. According to the thesis, other compression techniques such as distillation, model pruning, and reduced precision formats (fp16) appear to be on par or better in terms of the performance drop / compression rate. Generally, tensor decomposition methods can offer only a modest memory reduction without a significant performance gap. There are only a few examples where the proposed methods outperform other techniques for a comparable compression rate (e.g. WNLI for GPT-2 in Table 5.3). Thesis also lacks an analysis of the inference time for models modified with TTM or FWTTM FC layers. This is an important consideration, as tensor contractions require additional computational time and could introduce extra performance overhead. This overhead might be an additional factor to consider when selecting a memory reduction technique in practice.

Nevertheless, the thesis is a notable development of tensor decomposition methods for neural networks since it overcomes many previous obstacles. Obtained results show that tensor decomposition methods still require more investigation for wider application in neural models deployed in practice.

Presentation and the relevance of the topic of dissertation work to its actual content

The quality of presentations could be improved in several ways.

While most of the content is coherent and relevant, Chapter 6 seems to be disconnected from the topic of the thesis. It is related to an adjacent work on comparative question answering. There is a small part in the end related to compressing ColBERT – a ranking model for retrieved arguments, but most of this chapter discusses other aspects of comparative QA, including other ranking models, important features, and a shared task submission. I believe the thesis would benefit from making Chapter 6 more focused on the main topic related to the tensor decomposition, even if that requires removing some material and information related to the adjacent work.

Some chapters have clear artifacts, revealing that the content was taken directly from the author's publications. While the thesis should be indeed a compendium of publication ideas and results, such artifacts as repeating introductory motivations, contributions, overlapping related work sections, and background in various chapters make it hard to follow the main idea of the work. For example, introduction and related work in Chapter 7 seriously overlap with corresponding sections in Chapter 5. Overall, the thesis would benefit from clustering these pieces of information together. There are also other minor artifacts like reference to the future release of code on the page 73.

Some other minor suggestions regarding presentation are given in the "Other remarks" section of the review.

Publications

The results of the thesis are published in 5 articles (+ 2 articles under review). Among published works, one article presents results with MEKER, and 4 publications are related to comparative question answering.

Questions to the author

1. FWTTM-based models demonstrate substantial improvements over TTM-based Transformers on text generation and NLU tasks. Is it possible to apply FWTTM to GPT in the same evaluation setting on language modeling, NLU, and text summarization tasks?

2. GPT-2 small is comparable to Distill GPT-2 in terms of size, but its quality is not presented in Tables 5.2, 5.3 where you compare various compression techniques. Why do you not simply compare these techniques with an originally smaller model?

3. Is it possible to combine tensor decomposition methods with a reduced precision format (fp16) to obtain better compression?

4. What is the compute time overhead for inference of TTM-based models?

5. In table 7.1, you specify the size of various Transformer components. Attention is presented as a dedicated component. Nevertheless, self-attention consists of multiple FC layers. Have you tried to compress FC layers in attention?

6. Are there other works that successfully apply TTM for neural networks?

Other remarks regarding presentation

1. Tables 5.3, 5.4 could benefit from adding a comparison with other reduction techniques (e.g. DistilGPT).
2. Tables 3.1 and 3.2 could be combined. This would be possible if all values are multiplied by 100, and result and std values are presented with one digit precision.
3. There are several examples when acronyms and formula notations are given after they are used in text for the first time or not given at all. Consider FWTTM, sign B on the page 62, sign L in some formulas, F in 3.1, etc.
4. The capitalization of titles is quite inconsistent. You can choose to capitalize the first letter of all meaningful words or simply capitalize the first letter of the first word.
5. Figure 6-5 is incorrectly cropped and is not in a vector format.
6. Formula 2.5 referenced on the page 61 needs a short naming in addition to the reference.
7. There are some repeating phrases along the text, consider the page 119: "For SVD we use built-in Python function …"; "It is interesting to note …" on the page 33.

Overall assessment

Overall, the thesis presents high-quality research, which makes a notable contribution to the field of tensor decomposition methods for compression of neural networks.

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

[redacted]

☐ *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*