Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

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: Prof. Alexander Panchenko

The thesis document includes the following changes in answer to the external review process.

Response to Dr. Valentin Malykh

p. 27 "L2 norm between the and" – there is a word missing here.

This line is corrected to "This approach aims at minimizing L2 norm between the appropriate combination of factor matrices and corresponding unfolding"

Response to Prof. Pawan Goyal

In terms of writing, the thesis requires a thorough proof-reading. I am attaching some corrections on the thesis but these may not be exhaustive.

All corrections, highlighted in the additional material, are addressed. We are not able to hire a professional proofreader; instead we applied a spellchecker to the full text of the Thesis.

I had a question with Table 5.1. TTM is one way to reduce the number of parameters. A baseline to compare could have been to use GPT-2 small with similar number of parameters (by changing no. of attention heads / feed forward layer). Would the TTM based framework always be superior to actually reducing the number of parameters? This question may be applicable for other methods as well.

This is a very interesting question. While pruning or distillation can also be used to reduce the number of parameters.
However, for effective pruning, it is necessary to correctly select the layers / heads that we will prune (as, for example, it is done in or) - this requires additional research. Distillation, on the other hand, requires fine-grain selection of training hyper parameters. We did not have the resources to reproduce these approaches properly, and there were no models pre-trained using these approaches on the webText-103 dataset, so there is no such comparison in Table 5.1.

The pretrained models are for OpenWebText, so in Table 5.2 we were able to compare our method with Distill GPT-2 and with the OPT model - a model with a reduced number of heads and layers, trained from scratch on a strong mixture of datasets. The OPT gave a better result than our method, which is more due to the good quality of the training data. The Distill GPT did not perform very well.
Summarizing, it is worth saying that the quality obtained in the compression method depends not so much on the method, but on the data and the training pipeline.

A general suggestion would be to argue on the connection of the proposed method with the contemporary approaches such as LoRA. Also, I did not find any suggestions for future work in the last chapter. Maybe those can be added.

Thank you very much for this comment.
LoRA is a very strong approach and a good baseline, which reduces the number of parameters in the matrix of weight increments δW. It leads to decreasing the memory needed for training steps and does not affect the size of the initial model.
In the Thesis we focus on compressing the initial model, which leads to both decreasing memory needed for placement models in the system and decreasing memory needed for training.
By the way, on the point about memory reserving for training, we can use Lora as a baseline. Furthermore, it can be interesting to apply the TTM decomposition to δW instead of W and test the result.
A text about this has been added to the Thesis at the end of Chapter 6.

Also, all noted by Dr. Pawan Goyal grammar and technical misprints were addressed in the revised version of the Thesis.

Response to Prof. Steffen Eager

Firrst, there are several typos or ungrammatical statements (such as "intriduced", "it losses the possibility", "presiesely", "The matrix ia reshaped", "Than axis are permute", "Now cores store store only").

We applied spellchecker to the text of the Thesis and fixed the proposed misprint in the final versions.

Second, the approaches are sometimes quite technical, with a stronger focus on mathematical exposition, at the neglect of explaining certain concepts to the reader. For example, as far as I can see, the concept of knowledge graph embedding is not explained in Chapter 3.

We enriched this Chapter by brief explanation of Graph embeddings.

Third, while I appreciate the mathematical nature of the thesis, I must observe that the math is often presented in a way typical of engineers, i.e., with a lot of unexplained symbols whose meaning must be inferred out of context and sometimes using matlab notation.

We reread the mathematical background and formulas and added the missing definitions (for example, in Eqs. 2.1- 2.5, 2.8).

Fourth, when I look at the evaluation part, I wonder whether all comparisons are rigorous. To take chapter 5 as an example, I notice that Table 5.2 compares five models, Table 5.3 compares four models and Table 5.4 three models. As this is unexplained, it leaves the impression on the reader that baselines are removed in order to present more convincing results? While this may not be true, an explanation should be given.

Thank you for this note.
Indeed, it is better if all presented baseline models participate in all experiments. We managed to complete the NLU experiments on missing models and updated Table 5.3. The results are similar to language modelling experiments: the best score among compressed models belongs to OPT, the model with TTM layers performs quite worse, but better than SVD. DistillGPT2 has the worst score.

Finally, some of the chapters are also presented in a way that clearly prioritizes mathematical modeling and results presentation over a wider discussion of competitor techniques and implications of the results.

We have added more discussion of the various techniques. For example, in Chapter 5, we added a description of methods for reducing memory when training a model, the motivation to choose exactly the compression by an algebraic representation and comparison of the results of mentioned approaches.

I want to point out a final fun fact: as far as I can see, there is only one chapter that starts with a quote, namely, chapter 6.

We decided to remove this quote for the consistency of the entire text.

Finally, page 47 takes about "due to space constraints", but as far as I can see, the thesis has no space constraints.

This was a misprint, and we removed it.

Response to Prof. Andrey Kutuzov

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.

Tensor decompositions/representations decompose/represent data into factor matrices or sparsely interconnected small-scale low-order core tensors.
In the Thesis, we represent data in a NN layer as a Tensor Train Matrix form, and Knowledge Graph data as a Canonical Polyadic form. All these forms help to reduce the number of parameters according to the initial tensor.

In Section 1.2 formulation "Transformer-based language model by replacing some layers with Tensor structures" is not good, it is replaced by "Transformer-based language model by replacing some layers with Tensor Train Matrix structures", since Tensor Train Matrix is a way of tensor representation form.
We also add necessary clarifications in Introduction, Chapter 1, Chapter 3 and Conclusion; we will add this moment into the presentation.

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.

Indeed, each chapter is based on a publication or group of publications. To avoid confusion, we have added the title of the underlying work at the beginning of each chapter.

3. Language
I definitely recommend the defendant to actually run a spell checker on the full text of the thesis, and probably hire a professional proofreader

We passed the test through the spell checker and addressed comments from additional material.

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").

We added the definition of language modeling concept to 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.

The Language in Chapters 5, 6 is English, we highlight it in the text.

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?

We define the size of OpenWebText (8 mln of texts).
OpenWebText corpus was chosen because of its similarity to WebText, a database on which regular GPT-2 (baseline we compare with) was trained. Training the GPT-2 size model from scratch on our university resources took approximately 2 months; therefore, we did not have the opportunity to repeat the experiment in other languages, although it would be very interesting.
We also add to the text a notion of Wikitext-103 size, as well as mentioning that it should be considered as a sandbox dataset.

4. Some equations are numbered and some are not (even within one chapter). It does does not look good and does not help with navigating the text.

We made every equation numbered (for example in Chapter 2 and Chapter 4).

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.

We load models described in the Chapters 5, 6 to HuggingFace Model Hub (https://huggingface.co/s-nlp/) and add links into the Thesis.

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?

These people are bachelors from the University of Hamburg, aged 18-22, of different genders. The native languages are unfortunately unknown.

7. Table 6.10 should explicitly specify which sub-table represents scores for relevance/quality.

We fixed it in the text.

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 are on par with the rest. Is it just a bad choice of an example or…?

It is a bad choice of example. We changed it.

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.

Tables 7.5 and 7.6 each relate to a unique type of experiment (Single-train or Double-train).
Single Train is one experiment and consists of actions: fine-tuning and compression. Double-Train is one experiment and consists of actions: fine-tuning, compression, and further fine-tuning. In these tables, the columns differ in compression methods and compression ratios, not for the experiment type.

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).

We set Tables 7.7, 7.8 in the proper order.

Also, all noted by Prof. Andrey Kutuzov grammar misprints were addressed in the revised version of the Thesis.

Response to Prof. Alexey Frolov

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

In Table 4.2 the measurement units are megabytes
(and are returned by TORCH.CUDA.MEMORY_ALLOCATED).

I read once more the introductions to the Chapters and removed all the repetitive constructions.

Response to Dr. Artem Shelmanov

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?

Yes. TTM functionality, as well as the Fisher Weighted functionality, is integrated at the level of the Pytorch Fully Connected layers, so it can be applied to every Transformer model in the evaluation setting of the proper task.

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?

Yes, such an addition can be informative. We did not do it due to limited time.

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

Yes, TTM is compatible with fp16. Moreover, we provide several of these types of experiments on GPT-2 and RuDOLPH (multimodal Transformer architectures for Russian). The result will be added to the presentation.

4. What is the compute time overhead for inference of TTM-based models?
For example, for GPT2 there are
without TTM:

Averaged forward-backward time per batch: 0.0079 seconds
Average power comsumption for forward-backward pass per batch: 6.69e-07 kWh
Memory in forward-backward: 1518.64 MB

With TTM:

Averaged forward-backward time per batch: 0.0073 seconds
Average power comsumption for forward-backward pass per batch: 7.34e-07 kWh
Memory in forward-backward: 906.56298828125 MB

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?

Yes, we tried. The FC layer in the attention module is compressed with the same compression level as the normal one. Unfortunately, the quality of such a model (both trained from scratch and obtained by the decomposition of a pre-trained layer matrix) is much less than desired.
We explain this by the fact that the linear layer in Attention is, in fact, concatenated of Query, Key, and Value matrices. Creating a TTM representation from a given matrix requires shuffling elements, as described in Section 2.2.4. Thus, a new "matrix of independent block matrices" is created, over which the

"tensor of matrices" object will be created. This shuffling breaks the structure query, key, value in the initial matrices and breaks the Attention mechanism.

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

Not so much.

In 'Tensorized embedding layers for efficient model compression', Valentin Khrulkov, Oleksii Hrinchuk, Leyla Mirvakhabova, and Ivan V. Oseledets represent embedding layers in a TTM form.

We also addressed all remarks made by Dr. Shelmanov concerning the Thesis text.