Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Alexander Fonarev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Matrix Factorization Methods for Training Embeddings in Selected Machine Learning Problems

Supervisor: Prof. Ivan Oseledets

Chair of PhD defense Jury Prof. Andrzej Cichocki ***Email***: *a.cichocki@skoltech.ru*

Date of Thesis Defense: 19 September 2018

Name of the Reviewer: Dmitry I. Ignatov

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** **Date: 5-09-2018**

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

As advised, I structure my review by the sections below:

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

The thesis (121 pages in total) contains title, abstract, the list of published papers and delivered conference talks, acknowledgements, eight chapters including introduction, conclusion, and appendix with an auxiliary theorem. The structure is well balanced since it provides a necessary background material, the author's theoretical and technical contributions into three methodologically interrelated areas (addressing categorical features for machine learning applications, natural language processing by low-dimensional representations, and cold-start problem in recommender systems). The results chapter confirms the practical value of the proposed techniques. Each important chapter contains its own conclusion. Thus, this is the highest expected quality to my view given the size restrictions.

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

I found the title of the thesis as highly relevant to the content. In fact it invokes a trendy jargon term from deep learning community – embedding – and demonstrates how matrix factorizations can shed light on this mathematically obscure term.

- The relevance of the methods used in the dissertation

The author proposes a natural view of the so-called embeddings as results of low-rank matrix factorization techniques, which provide us with alternative exemplars of the phenomenon w.r.t. the existing examples of embeddings in deep learning community. Thus, the methods are highly relevant to the study providing their user with low-dimension vector representations of the original data objects as the embeddings intended to do as well.

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

All the results seems to be well-validated at the international level conferences by Data Mining, Natural Language Processing (NLP), and Information Retrieval communities. Two A* conferences are in the list of venues where the author of the thesis presented his joint works.

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

Being an expert in selected topics from Machine Learning, Recommender Systems, and Information Retrieval domains, I can confirm the relevance and importance of the proposed methods for feature generation in machine learning and solution of the cold start problem in recommender systems domain. As for NLP, the results obtained seem to be a successful solution for similarity-preserving matrix factorization for words representation.

- The quality of publications

The publications are at very good international level. Two of them are presented in the proceedings of two leading conferences on Data Mining and Natural Language Processing. The remaining ones are published in the peer-reviewed volumes, maintained by Springer in particular.

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

Overall, the command of English at the top level, but there are some language inaccuracies.

1. As it is illustrated on Figure 2, – in Fugure

2. where the embeddings are be used. – are being used.

3. close to each other in the representation space and, that is why a separate dataset which contains level of similarity between some of objects could be used as a benchmark to measure an embedding algorithm performance. – this piece should be definitely rewritten.

4. There are a large number of papers – the papers are in the focus

5. [106] explores using word embeddings

6. [58] explores items embeddings in recommender embeddings, – the last embeddings term should be systems, is not it?

7. when initial matrix A – when the original (or input) matrix A

8. using alternating approache

9. the gradient step an the retraction

10. After this, the algorithm – After that, or After this step.

11. Decision tree-based methods vs Decision tree based methods

12. Although single decision trees [15] usually do not show a good performance, and ensembling decision trees often provide state-of-the-art results. – Although single decision trees [15] usually do not show a good performance, while ensembling decision trees often provide state-of-the-art results.

13. In case of the binary classification – which one do you mean, the is a void identifier here.

14. so-called log-loss – the is missing

15. A neural network decision-making process consists of several steps (layers), and each of these steps takes outputs from the previous step (neurons), applies a non-linear transformation to them and multiplies them by some weight matrix in order to generate outputs. – The sentence should be rewritten since applies actually does not related to a proper noun word.

16. We get new transformed matrix Z – a is missing.

17. Step 1. Objects are texts.described

18. if matrix some loss-function ρ(·, ·) is used?

19. All results are statistically significantly different, which is proved by Mann-Whitney U test. – All the results…

20. Figure 14. Coverage or diversity. – Coverage and diversity

There are several terms that sounds slightly weird to me:

1. There are three main reasons to train embeddings – are those embeddings supervised machine learning techniques?

2. the interception of columns C and rows R – the term interception was not introduced

3. Each column of C contains the coefficients of the representation of a row in U via the vectors from S. – where this U has been introduced?

4. These features are called categorical, nominal or factor. – simply factors?

5. It means that many widely used and very powerful real-valued techniques, e.g., Random Forests [14], cannot be efficiently applied to these tasks. – It depends on input datasets and basic decision tree inducers within the ensemble.

6. In equation (30) usually alpha is out the brackets in the nominator, and alpha multiplied by the number of feature values in the denominator.

7. Let D be a multiset of all word-context pairs observed in the corpus. – Do you really use a multiset here?

8. The obtained factor P in Rn×r – Usually factors are columns or rows of the resulting product matrices.

9. The most standard way to

10. feature values co-occurrence frequencies, – may be co-occurrence frequencies of feature values

11. in eq. (41) and (45) the universal quantifier is used after the variables it relates.

12. In eq. (45), should the equality sign be by definition sign? Similarly, for eq. (47).

13. we have also experimented – Capital We.

14. i not in k and add it to the seed set: – it is hard to track where an index or vector is used without proper notation.

15. we use the 5-fold cross validation with respect to the set of users in all experiments. – why THE 5-fold?

Several questions for discussion.

1. Note that FM is a complete supervised prediction model that can handle categorial features, but it is not a categorical feature transformation method. – Categorical (typo). It is not, but the baseline predictors and pairwise factorized feature weights could be extracted and used as features as RecSys Challenge 2018 winners (3rd place) do, for example.

2. In case of explicit feedback, the most popular approach for sparse matrix factorization is Alternating Least Squares (ALS) method [10]. – Accordoing to Y.Koren, he suggests using ALS for implicit feedback.

3. This framework is a folklore knowledge [108], – I would rather agree, if you mean embeddings term.

4. Probabilistic Latent Semantic Analysis (PLSA) [46] solves the problem similar to LSA. The only difference is the loss function ρ used in the optimization. – this should be commented whether always the input matrices are stochastic or TF-IDF is used.

5. Step 1. Objects are values of categorical features. Each categorical feature’s value is described by a vector of co-occurrence frequencies with another categorical feature’s values. So each categorical feature’s value has as many embeddings as is the number of the rest categorical features in the dataset. This is a novel approach introduced in this work.

6. This might be a novelty, but there are well-kwnown continegency tables or Quetelet coefficients in Statistics similar to the features used in the algorithm.

7. Step 1. Objects are natural language words. Each word is described by a vector of co-occurrence frequencies of encountering this word within the same context with any other word from the dictionary.

8. What about the positions of such word, are they technically coincide?

9. It is described by the following categorical features: User ID, Item ID, User social information, Genre information. – but in Table 2 there are six features without their names.

Table 3. The parameters of the methods used are not provided. E.g. SVM kernel type and its parameters.

10. Moreover, the best results were obtained when SVD-SPPMI embeddings were used as an initialization of Riemannian optimization process. – How costly such initializing preprocessing

is?

10 minutes and 70 minutes respectively - Do the mentioned 70 mins include those 10 mins?

11. Figure 9. It seems RO-SGNS needs early stopping to prevent its overfitting.

12. Algorithm 9. Where T initialized at step 8 is used later?

13. As for the Rectangluar MaxVol algorithm,, this is an academic implementation of the idea used by such company as Imhonet in the past, showing that the idea is really useful.

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*