Skoltech

Skolkovo Institute of Science and Technology

From: Associate professor Victor Lempitsky
Date: August 29th, 2018

Thesis examiner report

For the PhD Thesis entitled "**Matrix Factorization Methods For Training Embeddings In Selected Machine Learning Problems**" by Mr. Alexander Fonarev.

Overall assessment

Learning vectorial embeddings of non-vectorial data is a fundamental task in natural language processing and information retrieval fields that has multiple practical applications. The presented thesis covers significant advances related to this task. The advances are motivated by the applications and are based on the matrix factorization methods. They include a new method for learning embeddings of categorical features, a new word embedding method based on Riemannian optimization, and a new approach for obtaining embeddings of so-called "cold" users and items in recommender systems based on the maxvol algorithm.

Overall, the thesis represents a solid work with clear contribution to state-of-the-art of the field of study. The main results have been presented in three publications, for which Alexander is the first author. Two of the three publications are top-level publications in natural language processing and information retrieval (the ACL conference and the ICDM conference). The thesis is well composed and well written.

Thesis overview

The thesis contains seven chapters and an appendix. The first chapter includes introduction and covers the topic of embedding learning as well as the important applications of this technology. It thus provides the motivation for the thesis.

The second chapter provides an overview of several topics related to the contributions of the thesis. This includes matrix factorization, supervised machine learning (including most popular classifier types), the use of categorical features in machine learning, standard

approaches to learning word embeddings, and basics of recommender systems and associated problems (including the cold start problem).

Chapter 3 is short and presents the general framework for the embedding learning using low-rank matrix factorizations, it also discusses how several popular approaches fit into this general framework. It further briefly discusses how the proposed new methods fit the same framework. Chapter 3 thus concludes the introductory part of the thesis.

Chapter 4 presents the first contribution of the thesis, which addresses the problem of learning real-valued embeddings for entities containing categorical features. The idea behind the new method is to map each object to a very high-dimensional vector based on the second order co-occurrence features. While the resulting vectors are embedded into Euclidean space, they are too high-dimensional for most practical purposes. Therefore this dimensionality is reduced using low-rank factorization. The quality of the resulting embeddings is evaluated on standard datasets. The evaluation is indirect, as the quality of the learned unsupervised embeddings is evaluated by measuring the success of a standard machine learning classifier (random forest) on the resulting embeddings. To the best of my knowledge this is the most meaningful and conventional way to evaluate unsupervised embeddings.

Chapter 5 presents the second contribution of the thesis. It observes that many problems in word embedding learning correspond to finding a low-rank matrix, where each column corresponds to the embeddings, and which minimizes a certain objective, which can be quite complex and problem specific. Often the task is handled in two steps, where a full-rank matrix is found first by minimizing the objective, and then a low-rank approximation is found in a separate step, which is guided by a simpler objective not necessarily related to the original objective. As this two-step process optimizing two different objectives often leads to suboptimal results, it is proposed to merge the two steps of optimizations into a single optimization process over the domain of low-rank matrix manifold (Riemannian optimization). The original two-step process may serve as an initialization. The experimental part of the section convincingly demonstrates the advantage of the joint optimization in terms of the resulting objective, as well as the task-specific measure (semantic similarity prediction).

Chapter 6 describes the third contribution of the thesis that addresses a common and important problem in recommender systems known as the “cold start” problem. When a new user (items) joins the recommender system, one needs to assign it a credible embedding based on the minimal amount of elicited information. Under the assumption that there is a limited number of representative items that the user can evaluate in the initial stage (a limited number of representative users that can evaluate the item), the new

method is proposed for the selection of such representative objects. The method is based on the generalization of the maximum volume submatrix (maxvol) algorithm, where the volume of a rectangular submatrix is defined as the product of singular values (which generalizes the common definition of the volume of square matrices through determinant). The selection of representative items then proceeds via greedy selection of items so that their embeddings maximize the rectangular volume. The chapter also discusses how the rectangular volumes of multiple submatrices considered by this greedy process can be computed efficiently. A complexity analysis and an experimental comparison with standard maxvol approach to item selection conclude the chapter.

Lastly, Chapter 7 contains a short overview and a concluding discussion of the results.

Conclusion

In my opinion, the presented manuscript **"**Matrix Factorization Methods For Training Embeddings In Selected Machine Learning Problems**"** by Mr. Alexander Fonarev constitutes an important contribution to the fields of embedding learning and recommender systems with wide practical applicability and good results. Overall, the presented thesis fully qualifies for the requirements of the PhD degree.

Yours Sincerely,

Victor Lempitsky

Associate professor at Skolkovo Institute of Science and Technology (Skoltech)
Samsung AI Center Moscow, Lab leader

lempitsky@skoltech.ru v.lempitsky@samsung.com
+7 916 391-47-73