POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A reproducibility study on Recommendation Systems of papers published in IJCAI and WWW international conferences

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Riccardo Luigi Aielli

Advisor: Prof. Maurizio Ferrari Dacrema

Academic year: 2022-2023

1. Introduction

In recent years recommender systems, have become very popular, this is probably due to the growing economic interest in online platforms like Amazon, Netflix, Instagram, Spotify, and many others. These systems try to anticipate user preferences or ratings for the available content on these platforms in order to suggest the right content to the right users avoiding inappropriate information overloading and enhancing platform profitability. In numerous recent research papers on recommender systems, there is a common assertion of substantial advancements compared to the existing "state-of-the-art". Nevertheless, there is no unanimous acknowledgment of this success, and various papers have raised concerns regarding the actual attainment of the claimed results. In this work, following the same approach used in previous studies [3], we carried out a comprehensive set of experiments to reproduce the results presented in recently published papers from leading scientific international conferences: International Joint Conference on Artificial Intelligence (IJCAI) and World Wide Web (WWW). We assessed the actual performance of newly proposed methods by comparing them against appropriately tuned well-established baselines using different evaluation metrics.

2. Motivation

In the realm of contemporary recommender systems research, authors have considerable flexibility in the selection of experimental conditions, this includes the choice of metrics, evaluation protocols, datasets, and baseline models. This leads the authors to present results where their proposed method performs optimally, often without providing a rationale for their choices. Moreover, what these papers often neglect to explicitly mention is that such claims are typically contingent on highly specific experimental conditions. These problems are emerging particularly in those algorithms integrating neural techniques such as Deep Learning, which are becoming extremely popular in this field.
Numerous studies [5] have also showcased the ability of older methods to outperform newer algorithms when they are appropriately fine-tuned. This phenomenon is often attributed to the absence of standardized benchmarks against which new algorithms can be effectively assessed.
Another significant issue related to reproducibility pertains to the common scenario where a paper introducing a new algorithm asserts to offer

a precise and comprehensive explanation of his work, but, in reality, falls short in providing adequate information about the technical methodology employed and the setup used for evaluation. Frequently, there are gaps in details, reliance on proprietary data, or disparities between the stated approach and the actual implementation. Additionally, the publication often does not include the link to the source code.
Since issues of reproducibility and competitiveness have already been pointed out in the past, we are now wondering if the situation has improved or not by conducting a similar study. Following the same approach used in [3] we are going to run the same experiment on other more recent papers taken from top leading conferences. To perform our experimental work we assess the new proposed algorithms within the same previously employed framework [4], which has unveiled substantial disparities between the reported and the validated outcomes.
Our primary objective in this work was to validate the feasibility of reproducing the reported results, ensuring the absence of methodological evaluation errors, and gauging the competitiveness against conventional non-neural approaches.

3. Paper Selection Criteria

In order to select the paper to conduct our experiment we followed a precise method. First of all, we distinguished three different categories of papers based on some constraint that needs to be respected. We defined as *relevant* papers all those articles that meet the following criteria:

- The paper had to be related to the recommender systems domain.
- The paper introduced a novel neural recommendation technique for top-K recommendation either pure Collaborative Filtering or a Hybrid method using precomputed item features. (e.g., methods that are trained on non-structured data such as images or reviews are not included, while methods that use pre-trained embeddings are included). Papers of the following categories are excluded: Cold Start, Sequence Aware, Review-Based, and News Recommendation.
- The paper had to be published between 2021 and 2023.
- The paper had to be published in the proceedings of either the International Joint Conference on Artificial Intelligence (IJCAI) or the World Wide Web (WWW) Conference, workshop papers are not included.

While we define as *candidate* paper the *relevant* papers in which we have all the needed artifacts to perform our reproducibility experiment:

- The source code must be available in its wholeness. If the source code is not publicly available we instantiated contact with all the authors asking for it.
- At least a dataset needs to be publicly available.

Lastly, the *reproducible* papers are the *candidates* ones in which we successfully confirmed the performance results declared in the paper.
We first processed the proceedings of IJCAI 2022 and 2021 (IJCAI 2023 had not yet taken place at the time of our selection phase) and then we moved on with the WWW 2023, 2022, and 2021 conferences until we got a reasonable amount of relevant papers to conduct our study. We meticulously reviewed all articles, looking for papers that were aligned with our relevant selection requirements. If the article did not include the source code required to reproduce its experiments, we initiated contact with all of the paper's authors to request access to it. If we did not receive a response within a month, we made a second attempt to contact the authors. If the second attempt was unsuccessful, we recorded the absence of a response.
Once we had acquired the source code, either from the article itself or through communication with the authors via email, we classified a paper as a *candidate* if it included the entire algorithm model, encompassing its training process, and the code necessary for generating item prediction and at least one of the datasets used for evaluation in the original paper along with comprehensive information concerning data preprocessing and partitioning.
Once we defined which were the *candidate* papers for reproducibility we started our reproducibility experiments porting examined algorithms into our framework following the reproducibility guidelines of the Association for Computing Machinery[1] (ACM).

4. Candidate Papers with Non-Executable Source Code

After having analyzed accurately more than 50 papers we found 9 *relevant* papers of which only 6 were considered as *candidates* for reproducibility according to our previously discussed criteria. For two of them was not possible to port the models successfully in our evaluation framework, Table 1 lists those papers.

Year	Title
2022	RecipeRec: A Heterogeneous Graph Learning Model for Recipe Recommendation [6]
2023	Multi-Behavior Recommendation with Cascading Graph Convolution Networks [2]

Table 1: List of papers classified as *candidates* for which we didn't manage to successfully execute the experiment.

In the first case *RecipeRec: A Heterogeneous Graph Learning Model for Recipe Recommendation* [6] the authors formalize the problem of recipe recommendation with graphs incorporating the collaborative signal into recipe recommendation through graph modeling proposing a novel heterogeneous graph learning model. Unfortunately, the source code was not clear in its structure. After two weeks of considerable unsuccessful efforts in porting the model into our framework, we realized that a complete refactoring of the code was needed. However, rewriting code is not compatible with a reproducibility study since this kind of work to be properly conducted involves the use of original artifacts as outlined in the guidelines of the ACM[1].
While in the second case *Multi-Behavior Recommendation with Cascading Graph Convolution Networks* [2] the authors propose a novel multi-behavior recommendation model with cascading graph convolution networks named MBCGCN. In MBCGCN, the embeddings learned from one behavior are used as the input features for the next behavior's embedding learning after a feature transformation operation. The original implementation leverages some libraries that are no longer supported (i.e., deprecated libraries). This prevented us from reproducing the algorithm. It is important to note that publishing source that uses outdated versions is a practice that needs to be avoided as it makes impossible to reproduce the experimental setup due to the incompatibility of libraries.

[1] `https://www.acm.org/publications/policies/artifact-review-and-badging-current`

5. Reproducibility Outcome

We then proceeded with the porting of the four remaining candidate papers. Table 2 shows which papers are classified as *reproducible* among those that were both executable and *candidates*.

Year	Title	Rep.
2023	Automated Self-Supervised Learning for Recommendation [8]	Yes
2023	Bootstrap Latent Representations for Multi-modal Recommendation [9]	Yes
2023	Multi-Modal Self-Supervised Learning for Recommendation [7]	Yes
2022	Fast Variational AutoEncoder with Inverted Multi-Index for Collaborative Filtering [1]	No

Table 2: List of executable papers along with their reproducibility outcomes according to our criteria.

We successfully managed to reproduce the experiment for the first three papers.
Automated Self-Supervised Learning for Recommendation [8] proposes a unified Automated Collaborative Filtering (AutoCF) to automatically perform data augmentation for recommendation. Focusing on a generative self-supervised learning framework with a learnable augmentation paradigm. In this experiment, we were able to reproduce the results with even better performance in two out of three datasets. A critical aspect is that some of our simple baselines easily overperform this method. This is, unfortunately, a common outcome present in the majority of the analyzed papers.
Bootstrap Latent Representations for Multi-modal Recommendation [9] studies the multi-modal recommendation problem, where the item multi-modality information (e.g., images and textual descriptions) is exploited to improve the recommendation accuracy. The authors propose a novel self-supervised multi-modal recommendation model (BM3), which requires neither augmentations from auxiliary graphs nor nega-

tive samples. During the analysis of this paper, we were able to reproduce the values on both the datasets used, unfortunately, those results were largely outperformed by two baselines.

Multi-Modal Self-Supervised Learning for Recommendation [7]. The online emergence of multi-modal sharing platforms (e.g., TikTok, Youtube) is forcing personalized recommender systems to incorporate various modalities (e.g., visual, textual, and acoustic) into the latent user representations. Inspired by the recent progress of self-supervised learning in alleviating label scarcity issues, the authors propose a new Multi-Modal Self-Supervised Learning (MMSSL) method. We were able to successfully reproduce the results for all three datasets utilized. For two of them, we got slightly better results concerning all baselines. This method also reveals good abilities to leverage a wider number of items compared to the baselines.

Instead, we were not able to reproduce the declared outcome of the *Fast Variational AutoEncoder with Inverted Multi-Index for Collaborative Filtering* [1] method. In this article, the authors proposed a fast Variational AutoEncoder (FastVAE) trying to accelerate the slow computation of the softmax which has linear cost in the number of items. We run the experiment over two datasets. The performance results we obtained were compared to the performance declared in the original work and we noted lower values with a negative discrepancy in the range of 8% with the first dataset and a negative discrepancy of 17% with the second dataset. These results are too low and far from the declared one therefore we had to mark this paper as non-reproducible. This important gap in performance could be due to some key steps such as how the split of the datasets was performed or how they carried out the cross-validation which are left undocumented in the paper.

6. Discussion of the Reproducibility Experiments

Three out of nine relevant papers could be reproduced leveraging the provided source code and publicly accessible datasets. Even when all the necessary information was available, it was never easy to reproduce the algorithms and work with other people's source code. It was almost always insufficiently commented and often disorganized in structure and key passages. Also reproducing the environment needed for each algorithm posed some challenges.

Our study shows that none of the neural approaches considered emerges as a competitive and robust algorithm demonstrating strong superior performance in the top-k recommendation task. In the examination of each paper, we documented several methodological issues that may contribute to elucidating the observed limited competitiveness of the neural methods under consideration in this study. Some of them are summarized in the following lines.

Most of the time authors do not give any theoretical or practical rationale for their decisions. Frequently, crucial assumptions were left unjustified or unmentioned, with no supporting evidence of their choices.

New algorithms are often compared against other emerging methods as baselines and not compared to well-established baselines appropriately tuned.

Reading the papers we recorded a lack of documentation regarding the data split and preprocessing procedures which are key aspects of this kind of algorithms.

Moreover, the criteria employed to determine the number of training epochs is usually not specified.

7. Conclusions

The outcomes of the experiments appear to affirm that, within the scope of the papers analyzed in this study (albeit a limited subset compared to the overall publications), there has been irrelevant substantial progress in recent years. During our reproducibility study, we found that many of the problems we highlighted had already been pointed out in previous reproducibility works.

When replicating the scenarios outlined in the original papers, we observed that in the majority of these instances, the newly proposed methods are outperformed by the simple baselines. Additionally, technical considerations such as scalability, training time, recommendation time, complexity, and robustness which represent crucial aspects that can heavily impact user satisfaction, as well as development and maintenance costs, are most of the time

ignored.
To address these challenges and to address the methodological issues in the evaluation process outlined in this work, the recommender systems research community should adopt a more systematic approach to evaluating progress when introducing new technologies.
Enhancing the reproducibility of published articles stands out as a potential strategy to alleviate some of the identified issues.
Another open issue is also to understand whether the datasets we have are truly representative of reality or not. If there are industries such as Netflix that use neural methods massively but we can never find an advantage with public datasets it means that we have usage scenarios that the research community struggles to work on because of inadequate data. Consequently trying to get the most appropriate data should be a priority.

References

[1] Jin Chen, Defu Lian, Binbin Jin, Xu Huang, Kai Zheng, and Enhong Chen. Fast variational autoencoder with inverted multi-index for collaborative filtering. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 1944–1954. ACM, 2022.

[2] Zhiyong Cheng, Sai Han, Fan Liu, Lei Zhu, Zan Gao, and Yuxin Peng. Multi-behavior recommendation with cascading graph convolution networks. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1181–1189. ACM, 2023.

[3] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Trans. Inf. Syst.*, 39(2):20:1–20:49, 2021.

[4] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 101–109. ACM, 2019.

[5] Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. *CoRR*, abs/1905.01395, 2019.

[6] Yijun Tian, Chuxu Zhang, Zhichun Guo, Chao Huang, Ronald A. Metoyer, and Nitesh V. Chawla. Reciperec: A heterogeneous graph learning model for recipe recommendation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 3466–3472. ijcai.org, 2022.

[7] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 790–800. ACM, 2023.

[8] Lianghao Xia, Chao Huang, Chunzhen Huang, Kangyi Lin, Tao Yu, and Ben Kao. Automated self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 992–1002. ACM, 2023.

[9] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 845–854. ACM, 2023.