Harmonic attentive multimodal neural network for movie recommendation

J X Wang¹, Y Guo¹, ², ³, *, T M Qi¹ and Q F Tang¹

¹Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
²Business Intelligence and Visualization Research Center, National Engineering Laboratory for Big Data Distribution and Exchange Technologies, Shanghai 200436, China
³Shanghai Engineering Research Center of Big Data & Internet Audience, Shanghai 200072, China
*Corresponding Author: guoyi@ecust.edu.cn

Abstract. A multimodal neural network based on harmonic self-attention was proposed for movie recommendation. This method can deal with the multi-source data and learn representations of users and items well. The multimodal neural network mainly consists of three sub-networks, ResNet, Bert, and LSTM. In addition, the harmonic self-attention mechanism can explore the user’s preference from the perspective of time. Experimental results show that compared with other three latest recommendation methods, HSMNN shows better performance in terms of HR and NDCG.

1. Introduction

People living in the era of big data face a variety of videos, books, and news every day. The inexhaustible items make it difficult for people to accurately choose the one they like. People want to quickly find the items that suit their tastes, instead of wasting time in searching. The recommendation system helps people quickly and efficiently locate items that users are interested in.

Traditional recommendation algorithms can be divided into collaborative filtering recommendation algorithms (CF), content-based recommendation algorithms, and hybrid recommendation algorithms. The former two algorithms have their own shortcomings. Hybrid recommendation algorithm can make up for the shortcomings of a single recommendation algorithm and improve the performance of recommendation system. Based on the existing hybrid recommendation algorithms, this paper propose a multimodal neural network to represent feature of users and items. After that, a harmonic self-attention mechanism is designed with the time factor.

The main contributions of this paper are as follows:
- This paper proposes and designs a multimodal neural network with a harmonic self-attention mechanism to learn the feature representation of users and items.
- A harmonic self-attention mechanism is designed. It is used to provide higher attention weight to historical items that make more contribution based on user-item interactions. Besides, the time factor is added to the traditional self-attention mechanism.
- The popular movie recommendation dataset is adopted to experiments, and this dataset verifies the effectiveness of the proposed method.
2. Literature Review

The sparsity of interactions between users and items and cold start problems limit the performance of CF. Content-based recommendation algorithms usually extract the relationship between users and items by artificially designed features which not only limits the effectiveness and scalability of the system, but also ignores the interrelationships between the items.

In recent years, deep learning has made breakthroughs in image processing [1], natural language processing (NLP) [2] and other fields. On the one hand, the deep learning network is a kind of deep nonlinear network, which can better extract the deep features of the data. On the other hand, it maps multiple heterogeneous data to the same hidden space, and then obtains a unified data representation. [3] proposed a deep knowledge-aware network, which is a content-based deep recommendation framework and incorporates knowledge graph representation into news recommendation for click-through rate prediction. Experiments prove that DKN achieves greater benefits. HARSAM [4] divides the user’s potential feedback data with different time granularities, and learns the user’s potential preferences through deep neural networks.

3. Harmonic attentive multimodal neural network

3.1. Workflow of HSMNN

The workflow of HSMNN is shown in figure 1. It includes two modules, multimodal neural network and harmonic self-attention mechanism. The pre-processed features are input into the multimodal neural network with three sub-networks which can represent the characteristics of items well. The harmonic self-attention mechanism is a traditional self-attention network which added a time factor.

3.2. Multimodal neural network

The heterogeneous information in this paper can be divided into structured data, text data and image data. ResNet, Bert, and LSTM which included in multimodal neural network are adopted to learn and represent the image data, the text data and the structured data, respectively.

- Sub-network: ResNet [10]. Convolution-based deep learning networks can maintain the image's neighborhood relationships and spatial locality in potentially higher-level feature representations for image data. ResNet solves the problem that the deeper network stacking effect becomes worse when the depth of the network increases to a certain extent by adding residual network.

- Sub-network: Bert [5]. The NLP task is divided into two parts, word embedding by pre-training and downstream specific tasks. Google announced the performance of BERT in 11 NLP tasks in 2018.
It completely changed the relationship between the two parts in NLP task. Bert is adopted as a word embedding tool for downstream recommendation task.

- Sub-network: LSTM. Since time information and certain sequence relationship between the items are included in the structured data, LSTM is considered to be suitable for this kind of task.

3.3. Harmonic self-attention mechanism

All the historical items are treated equally may limit the performance of representation. Considering that the user’s interest may change over time, time factor should be taken into account when assigning the personalized attention weight. The time factor added to the self-attention mechanism is proposed, the equation is as follows:

\[
 w = \frac{\exp(f(p_i, q_j))}{\sum_{j \in R_u^+} \exp(f(p_i, q_j))} + (1 - \varepsilon) \frac{T_{j+1} - T_0}{T_j - T_0}
\]

where \( p_i \) denotes the vector of item \( i \), \( R_u^+ \) denotes the set of items that user has interacted with.

4. Experiments

4.1. Datasets description and evaluation schema

Each interaction is paired with 4 negative instances during training to ensure that the number of training instances is greater than the number of interactions. Leave-one-out evaluation protocol is adopted as performance evaluation, which uses each user’s latest interaction as test data and uses the remaining interactions for training.

Hit ratio (HR) and normalized discounted cumulative gain (NDCG) have been widely used for top-n recommendations. If their values are larger, the performance of the recommendation system is better.

\[
 HR@K = \frac{|H|}{|GT|} \tag{2}
\]

\[
 NDCG@K = Z_e \frac{2^e - 1}{\log_2(e + 1)} \tag{3}
\]

where \(|GT|\) denotes the number of groundtruth, \(|H|\) denotes the number of the positive instance in the top-k list, \( K \) denotes the number or top-k list.

4.2. Baselines and hyper-parameters setting

- NAIS [6]: NAIS is an item-based CF. The key component is an attention network that can distinguish which historical items in the user profile are more important for prediction.
- NAIRS [7]: Proposed a self-attention network as a key component aiming to assign attention weight to user historical items.
- NASM [8]: NASM is an attentive similarity model which captures both local and global item information by local attention and nonlinearly attention.

According to equation (1), it can be seen that the attention weight is directly affected by \( \varepsilon \). The value of \( \varepsilon \) is taken from \([0, 1]\) in steps of 0.1. Figure 2 shows that the two evaluation schemes all rise first and then fall when the value of \( \varepsilon \) becomes larger, which indicates that the best value of \( \varepsilon \) is 0.5.

4.3. Analysis of experimental results

The experimental results are shown in figure 3. In general, compared with NAIS, NASM, and NAIRS, HSMNN improves HR by 8.7%, 7.1%, and 4.1%, and improves NDCG by 12.2%, 7.0%, and 4.5%. NAIS does not consider the multi-source data, only adopts the movie ID; NAIRS adds a self-attention mechanism, but does not consider the time factor; NASM adds a local attention which obviously take time factor into account, but the local attention is linear and has poor performance.
5. Conclusion

HSMNN is proposed in this paper. It is a multimodal network with a harmonic self-attention mechanism for movie recommendation with multi-source data. It mainly solves the following problems: 1) A single neural network has poor performance of learning representation of the item and the user, the multimodal neural network with three different sub-networks is proposed for the image data, text data and structured data. 2) In order to assign personalized weights according to the different contributions of user’s historical items, a harmonic self-attention mechanism with time factor is proposed. It mainly adjusts the attention weights for the contribution degree of the user’s historical items and the effect of time. Finally, experiments on the MovieLens-1M are compared with other three state-of-the-art baselines and the experimental results show that proposed method outperforms.

Acknowledgments

This research is financially supported by The National Key Research and Development Program of China (grant number 2018YFC0807105), National Natural Science Foundation of China (grant number 61462073) and Science and Technology Committee of Shanghai Municipality (STCSM) (under grant numbers 17DZ1101003, 18511106602 and 18DZ2252300).

References

[1] M J Zhang and C Stefano 2018 Proc. GECCO 1221-57
[2] Ferrone, Lorenzo, and F M Zanzotto 2017 "Symbolic, Distributed and Distributional Representations for Natural Language Processing in the Era of Deep Learning: a Survey"
[3] H W Wang, F Z Zhang, X Xie and M Y Guo 2018 Preprint arXiv:1801.08284
[4] D L Peng, W W Yuan and C Liu 2019 IEEE Access 12620-29
[5] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser and I Polosukhin 2017 NIPS 5999-6009
[6] X N He, Z K He, J K Song, Z G Liu, Y G Jiang and T S Chua 2018 IEEE Transactions on Knowledge and Data Engineering 12 2354-66
[7] S Yu, Y B Wang, M Yang, B C Li, Q Qu and J L Shen 2019 Proc. WSDM 786-9
[8] Z P Shan, Y Q Lei, D F Zhang and J Zhou 2019 IEEE Access 7 70689-700
[9] F M Harper and J A Konstan 2015 ACM TIIS 5 NO.19
[10] K M He, X Y Zhang, S Q Ren and J Sun 2017 IEEE Conf. CVPR (Las Vegas: IEEE Computer Society) 770-8