2019 IEEE International Conference on Multimedia and Expo (ICME)

SEQUENTIAL BEHAVIOR MODELING FOR NEXT MICRO-VIDEO RECOMMENDATION WITH COLLABORATIVE TRANSFORMER

Shang Liu, Zhenzhong Chen[*]

School of Remote Sensing and Information Engineering, Wuhan University, China
shangliu@whu.edu.cn, zzchen@whu.edu.cn

ABSTRACT

Micro-video recommendation is important for micro-video social platform, which provides its users with micro-videos they may interested in. In this paper, we propose a new variant of Transformer, alleviating the drawbacks of Recurrent Neural Networks (RNN) which tend to compress all history records in a fixed hidden representation, to model users' sequential behavior for next micro-video recommendation. Firstly, we employ self-attention to capture micro-video's multi-modal features of different importance. Secondly, we make use of multi-head attention to learn users' preference from historical records. We show how the Transformer combined with Collaborative Filtering (CF) and user-video sequential interaction can be used to perform next micro-video recommendation. Extensive experiments are conducted on two collected micro-video datasets, i.e., Toffee and TikTok. The experimental results demonstrate the proposed method is more effective compared with several state-of-the-art sequential recommendation methods.

***Index Terms*— ** Recommendation, Micro-video, Multi-modal, Transformer, Collaborative Filtering

1. INTRODUCTION

Recent years have witnessed a spurt of progress in micro-videos social network, such as Musical.ly, TikTok, Toffee and so on. These platforms have attracted many users to share their daily stories through micro-videos and watch interesting micro-videos of others. With the user's online 'footprints' (user-video interaction) history, the micro-video social platform needs to predict next micro-video that the user may prefer and recommend it to him/her, which is of great value for improving user experience and app retention.

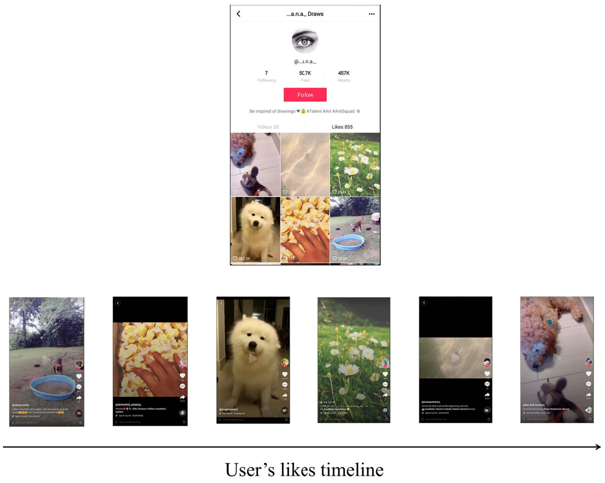


Fig. 1: An example of a user's likes timeline of micro-videos.

Micro-videos, which are user-generated with simple topics and extensive audience segments, gain popularity in a different way from traditional videos. Compared to traditional video recommendation, micro-video interaction sequence has a closer internal relationship and more stable preference pattern. Users usually watch relevant micro-videos in a short period, so the user's sequential behavior has a large impact on next micro-video recommendation. For example, Figure 1 shows a sample of user's board of likes which contains the micro-videos that the user has sent thumbs-up to.

Existing methods have achieved promising results in making sequential recommendation with user historical records. For example, Rendle et al. [1] proposed Markov chain to model user behavior sequences, and recently RNN has been employed to embed previously interacted items for current interest prediction [2]. However, RNN based methods tend to compress all of a user's history records into a fixed hidden representation, which cannot discriminate different historical records for next interest prediction. Recently, to alleviate the drawbacks of RNN, Transformer [3] was proposed to implement a stable multi-head attention mechanism and has demonstrated its superiority in machine translation against RNN based models.

In this work, we consider modeling user sequential behavior for next micro-video recommendation with a new variant of Transformer. Specifically, we leverage *self-attention* as the micro-video encoder for multi-modal information fusion.

[*] The corresponding author.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00086

Then, we feed the historical micro-videos' representations to a user preference decoder for sequential behavior modeling. Finally, the insights of Collaborative Filtering (CF) are incorporated in the score decoder to output preference probability for a new micro-video.

To this end, in this paper, we proposed a novel Transformer variant, named Collaborative Transformer, for next micro-video recommendation. Our main contributions are listed as follows:

- We propose to integrate the insights of CF with Transformer for next micro-video recommendation, which leverages multi-modal micro-video content information and user historical records in a more effective manner.
- A unified framework is built to subsume multimodal micro-video information fusion and user's sequential behavior modeling, thus the model can capture both micro-video representation and sequential user-video historical interactions.
- We constructed two real-world micro-video datasets. Extensive experimental results on these datasets verify the superiority of the proposed model in comparison with other state-of-the-art sequential recommendation methods.

2. RELATED WORK

Micro-video recommendation. Different from traditional online video recommendation [4], we focus on the recommendation of micro-video in social network, a new form of user-generated content, which is set as 15 seconds on TikTok and Musical.ly. Micro-video recommendation has attracted increasing attention from researchers [5, 6]. Chen et al. [5] proposed a hierarchical attention at item- and category-level for micro-video click-through prediction. Liu et al. [6] proposed a genre-aware micro-video recommendation model based on neural network.

Sequential recommendation. Sequential models usually rely on Markov Chains (MC) to capture user's sequential behavior patterns [1]. Rendle et al. [1] proposed to combine Matrix Factorization (MF) and MC for next-basket recommendation. There are also works that have adopted RNN for the sequential recommendation task [2, 7]. Dong et al. [7] proposed to combine RNN and MF in a multi-task learning framework capturing the whole historical sequential information. In this paper, we propose a new variant of Transformer combined with the insights of CF for next micro-video recommendation.

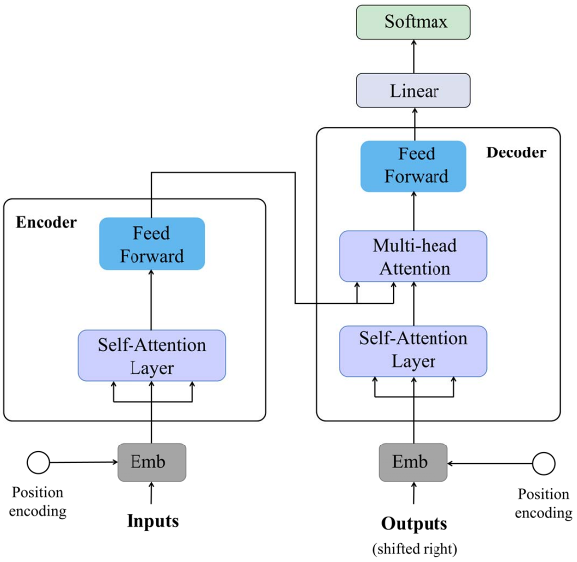


Fig. 2: Transformer with 1-layer encoder and 1-layer decoder.

3. PRELIMINARY

3.1. Notation and Problem Formulation

Before describing the details of our proposed model, we first introduce the notation and define the problem. Let $\mathcal{U} = \{u_1, u_2, ..., u_{|\mathcal{U}|}\}$ be a set of users, and $\mathcal{V} = \{v_1, v_2, ..., v_{|\mathcal{V}|}\}$ be a set of micro-videos. For each user $u \in \mathcal{U}$, we have a sequence of history micro-videos $H^u = (v_1^u, v_2^u, ..., v_{|H^u|}^u)$ that u has interacted with, where v_i^u means the i-th micro-video which the user u sends thumbs-up to. The sequential micro-video recommendation task we are tackling is to recommend a list of unseen micro-videos for a user to see next. Specifically, a higher prediction score of a user u to an unseen micro-video v_j indicates a higher probability that the user u will like it.

3.2. Transformer

In this subsection we introduce some background on Transformer [3]. Transformer is based on Encoder-Decoder, which both are composed of stack of identical layers. Figure 2 shows a Transformer with 1-layer encoder and 1-layer decoder. The way the attention mechanism is applied and customized is what makes the Transformer novel. Attention is a function that maps the 2-element input (*query*, *key-value* pairs) to an output. The output given by the mapping function is a weighted sum of the *values*, where weights for each *value* measures how much each input *key* interacts with the *query*. While the attention is a goal for many researches, the novelty of transformer attention is that it is *multi-head attention* and *self-attention*.

The *scaled dot-product attention* is the foundation of *multi-head attention* and *self-attention*. It is assumed that *queries* and *keys* are of d dimension and *values* are of d dimension. The input is represented by three matrices: queries' matrix Q, keys' matrix K and values' matrix V. The output of *scaled dot-product attention* is:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

The *multi-head attention* is composed of H paralleled *scaled dot-product attention* layers called 'head' which are independent with each other. The output of *multi-head attention* is:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^0 \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $W^0 \in \mathbb{R}^{d \times d}$, and $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{\frac{d}{h} \times d}$ are the independent head projection matrices, $i = 1, 2, ..., h$.

Multi-head attention imitates the classical attention mechanism where in encoder-decoder attention layers *queries* are hidden state from decoder layer, and *keys* and *values* are hidden state from the encoder. The *self-attention* is another case of *multi-head attention* where the *queries*, *keys* and *values* are all from the same hidden layer.

4. COLLABORATIVE TRANSFORMER

Our proposed Collaborative Transformer is illustrated in Figure 3. It consists of three parts. Micro-videos encoder convert a sequence of micro-videos into their representation, which discriminates different multi-modal feature embeddings with self-attention mechanism. User preference decoder works on the sequence to discriminate different historical records for the user with multi-head attention mechanism. Score decoder takes a new micro-video as input to predict its preferred probability for the user. In the following subsections, we will discuss each component of our model in detail.

4.1. Micro-videos Encoder

We seek an effective manner to convert each item into its representation, which discriminates different multi-modal features, such as unique ID m_i^{id}, visual content m_i^v, meta-information m_i^{mi} and position encoding m_i^p.

In this paper, the ID can be denoted as a one-hot vector m_i^{id} for m_i. We also train a linear embedding for it,

$$e_i^{id} = E_{id} m_i^{id}, \tag{4}$$

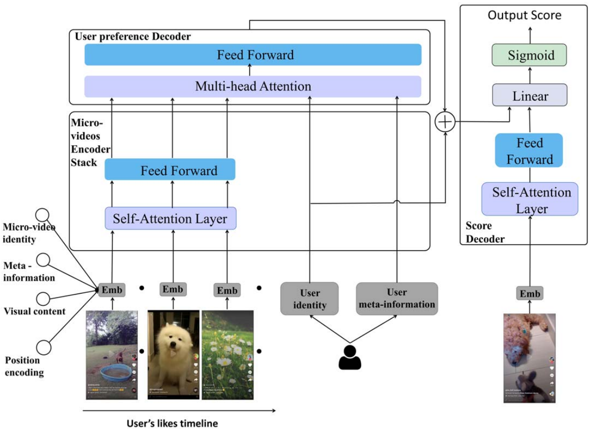


Fig. 3: The illustration of Collaborative Transformer framework.

where $e_i^{id} \in \mathbb{R}^d$ is an embedding vector for micro-video ID. m_i^v is produced by using an Inception-v3 model [1] pretrained on ImageNet, with image features at 1-frame-per-second extracted and average pooling on frame-level features. As m_i^v is high-dimensional, we perform a linear embedding for it,

$$e_i^v = E_v m_i^v, \tag{5}$$

where e_i^v is an embedding vector for micro-video visual content. m_i^{mi} is extract from micro-video meta-information such as author and background music ID. With the same embedding method as micro-video ID, we can get meta-information embedding e_i^{mi}. Since user's interaction with micro-videos is an ordered sequence, we use 'position encoding' m_i^p to represent the interaction order information. And we embed it in the same space as other features to get e_i^p.

Our Micro-videos encoder uses self-attention mechanism to discriminate these feature embeddings. The intensity of each feature may be affected by other features, which means the attention to some features may be weakened while others enhanced. The output of this layer can be seen as a fusion of micro-video multi-modal features to convert each item to its representation. We first concatenate these feature embeddings on the first dimension to a representation $F = concat^0\{e_i^{id}, e_i^v, e_i^{mi}, e_i^p\}$. It is then fed forward to micro-video encoder layer to learn a representation $F^e = E(F)$:

$$E(F) = \Phi(\text{FFN}(\Gamma(F)), \Gamma(F)) \tag{6}$$

$$\Gamma(F) = \begin{pmatrix} \Phi(MultiHead(e_i^{id}, F, F), e_i^{id})^T \\ \Phi(MultiHead(e_i^{mi}, F, F), e_i^{mi})^T \\ \Phi(MultiHead(e_i^v, F, F), e_i^v)^T \\ \Phi(MultiHead(e_i^p, F, F), e_i^p)^T \end{pmatrix} \tag{7}$$

[1] https://www.kaggle.com/google-brain/inception-v3

$$\Phi(\alpha, \beta) = \text{LayerNorm}(\alpha + \beta) \tag{8}$$

$$\text{FFN}(\gamma) = W_2 \text{max}(0, W_1\gamma + b_1) + b_2 \tag{9}$$

where $\Phi(\cdot)$ represents the function that performs layer normalization on the residual output, $\Gamma(\cdot)$ denotes the self-attention model, $\text{FFN}(\cdot)$ is the 2-layered feed-forward neural network, W_1, W_2 are the weights for feed-forward layers, and b_1, b_2 are the biases.

4.2. User Preference Decoder

User's historical accessed micro-videos can be plenty. If we use RNN to analyze the sequence $\{v_1^u, v_2^u, ..., v_{|H^u|}^u\}$, we may fail to capture both short-term and long-term dependency in it. Thus, we leverage multi-head attention mechanism in our user preference decoder, which discriminate different micro-videos in the historical records. First, we embed the user's features (e.g., user ID and user meta-information) in the similar way as that of micro-video. After getting user's embedding e_u^{id} and e_u^{mi}. We concatenate them on the first dimension: $e_u = Concat^0(e_u^{id}, e_u^{mi})$. With the historical sequential micro-video representation $F^e = \{F_{v_1}^e, F_{v_2}^e, ..., F_{v_{|H^u|}}^e\}$, we can perform the preference decoder as follows:

$$P = \Phi(\text{FFN}(\Omega(U, F^e)), \Omega(U, F^e)) \tag{10}$$

$$\Omega(U, F^e) = \Phi(MultiHead(U, F^e, F^e), U)^T \tag{11}$$

where $\Omega(\cdot)$ represent the multi-head attention model, $\Phi(\cdot)$ and $\text{FFN}(\cdot)$ is similar with Equation (6) and Equation (7).

4.3. Score Decoder

With the generated user preference decoder representation P_u, we first concatenate it with user identity embedding, $U_u = Concat^0(P_u, e_u^{id})$, to collaboratively predict user's preference on a new micro-video. The score decoder predicts the preferred probability of a new micro-video by taking U_u and new micro-video encoding F_v^e as input. The final loss function is the sigmoid cross entropy loss:

$$\sum_{u,v} -y_{uv} log(\delta(f(U_u, F_v^e))) - (1-y_{uv}) log(1-\delta(f(U_u, F_v^e))) \tag{12}$$

where y_{uv} is defined from the user's feedback, $y_{uv} = 1$ indicates that the user u likes micro-video v and $y_{uv} = 0$ indicates that there is no observed data about the interaction. f is a ranking function whose input is the micro-video encoding F_v^e and user representation U_u, which can be either a dot-product function or a more complexed deep neural network.

We uniformly sample negative instances from unobserved interaction according to the fixed sampling ratio of 1:1 to the number of observed interactions. An Adaptive Moment Estimation (Adam) algorithm is adopted to optimize the loss function.

5. EXPERIMENT

In this section, we conduct experiments on two collected micro-video datasets to compare the performance of our proposed Collaborative Transformer against other state-of-the-art sequential recommendation methods.

Table 1: Statistics of the datasets

Feature	#User	#Micro-video	#Feedbacks	Sparsity
Toffee	3,582	29,949	126,186	0.307%
TikTok	17,584	56,383	709,281	0.072%

5.1. Dataset

We constructed two datasets using micro-videos collected from Toffee and TikTok. To our best knowledge, there is no publicly available micro-video social media dataset that contains user-to-video preference interaction. Toffee and TikTok allow users to publish and share micro-video to others and send thumbs-up to preferred micro-videos. In this research, we crawled user's liked micro-videos from the 'likes' board of the user and the sequence sorting which indicates the order of adding to the board. To model user's sequential behavior, we filtered users with less than 5 user-video interactions and crawled the user's at most 100 sequential historical records. The statistics of these two datasets are listed in Table 1.

5.2. Comparison Methods

We compare our proposed Collaborative Transformer with a series of state-of-the-art baseline algorithms as follows: **BPR-MF [8]**, **RNN [9]**, **RNN+Attention [10]**, and **ATRank [11]**. We implement a attention-based RNN network for sequential recommendation as [10]. ATRank is a very state-of-the-art method on modeling user behavior as a sequence, which features a specially designed attention mechanism.

5.3. Evaluation Methodology

For each dataset, we take each user's last interaction record for the test set, the remaining as training set. The splitting method is widely used in previous work on sequential recommender system (e.g., [11]). Since it is too time-consuming to rank all items for every user during evaluation, we follow the common strategy [12], for each (u, v) pair in the test set, randomly sampling 500 additional micro-videos that are not interacted by the user and predicting scores by u for v and

Table 2: Recommendation performance in terms of HR@20 and NDCG@20

Dataset		(a) BPR-MF	(b) RNN	(c) RNN+Attention	(d) Atrank	(e) Collaborative Transformer	Improvement (e) vs. Best
Toffee	HR@20	0.3293	0.3516	0.3654	0.3884	**0.4618**	18.91%
	NDCG@20	0.1560	0.1642	0.1758	0.1704	**0.2044**	16.28%
TikTok	HR@20	0.3488	0.4028	0.4675	0.4460	**0.5238**	12.04%
	NDCG@20	0.1682	0.1921	0.2196	0.1860	**0.2313**	5.36%

the other 500 micro-videos. The evaluation criteria is to measure how well the method ranks the correct pair (u, v) against the other random micro-videos. We report the performance of each method on the test set in terms of the following ranking metrics: Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDCG).

5.4. Parameter Settings

Since the last hidden layer of models determines the model's capability, we term it as predictive factors and test the dimension size of the factors for all methods in the range of [8, 16, 32, 64, 128]. The batch size is set to be 128 for all methods. The regularization weights for l2-loss is set to be 5e-5. We use Adam as the optimizer. For the Collaborative Transformer, we set number of heads (h) to be 4 and the ranking function f is simply set to be the dot product in this task. If not specified, we show the results of predictive factors' size to be 64 and rank position to be 20 for all methods because of the consistent for comparison.

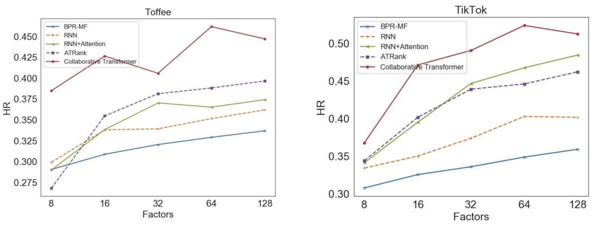
5.5. Performance and Analysis

The performance of all competitors with $\mathrm{HR@20}$ and $\mathrm{NDCG@20}$ on the two micro-video datasets is shown in Table 2. From the table, we have the following observation:

1) Firstly, we can see that BRP-MF has the worst performances because it only uses user-item interaction and does not consider the sequential user behavior. RNN models the user's sequential interaction on micro-videos for user representation and clearly outperforms BPR-MF, which proves the effectiveness of sequential behavior modeling. RNN+Attention outperforms RNN slightly since it integrate attention mechanism with learning of hidden state. ATRank considers the temporal behavior encoder and use self-attention to learn user representation from the history sequential interaction, which achieves a comparable performance to RNN+Attention.

2)Secondly, we observe that our proposed Collaborative Transformer is the top performing model on the two micro-video datasets in general. Collaborative Transformer outperforms RNN, RNN+Attention and ATRank, which are all recent competitive sequential recommendation methods. On average, our method outperforms the second-best method by 15.48% in terms of HR@20 and 10.82% in terms of NDCG@20 across the two datasets. We speculate that the proposed model integrate the insight of CF with Transformer for modeling user sequential behavior.

3) Finally, we can see that our proposed Collaborative Transformer has a good scalability, which performs well on both sparse and dense micro-video datasets. Compared to other methods, our model maintains the relative performance improvements across the two datasets Toffee and TikTok in general.



(a) Toffee — HR

(b) TikTok — HR

Fig. 4: Performance of HR w.r.t dimensionality of predictive factors on two datasets

Effects of number of predictive factors We compare the performance of HR with respect to different number of latent factors on the two datasets. The larger the size, the more expressive the feature can be, potentially leading to better performance. We only show the results of HR and the results of NDCG admit the same trend thus they are omitted. In Figure 4, we can see that increasing the factor size above 64 gives little gain. we choose 64 dimensions as a good compromise between performance and computational cost. On the other hand, our proposed model achieves a high performance with a small latent factor of 8, which indicates the high expressiveness of the proposed model by integrating Collaborative Filtering with Transformer.

Convergence analysis We proceed by demonstrating comparison of convergence rates of all methods on the two micro-video datasets. Figure 5 presents how HR@20 score improves with the training. The learning rate is set as 1.0 and epoch as 10. As we can see from the figure, the convergence ef-

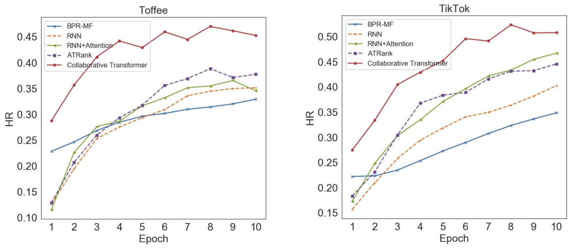


Fig. 5: Training curves comparison of Collaborative Transformer and baselines on Toffee and TikTok datasets.

ficiency of Collaborative Transformer is superior to all baselines. We can also see that our proposed model achieves much better performance than other baselines with the epoch, which again proves the effectiveness of the proposed model for next micro-video recommendation.

6. CONCLUSION

In this paper, we propose a new variant of Transformer combined with the insight of CF, called Collaborative Transformer, for next micro-video recommendation. Collaborative Transformer subsumes multi-modal micro-video features fusion and user's sequential behavior modeling in a unified learning framework. Specifically, we use a self-attention mechanism to discriminate different multi-modal micro-video's features, and customize multi-head attention for the user's sequential preference learning. Experimental results on two collected micro-video datasets demonstrate the superiority of our model in comparison with state-of-the-art methods, which show that our method is able to leverage user historical records more effectively. In the future, we would consider incorporating more context information such as social network and textual comments content into the model to further improve the sequential micro-video recommendation accuracy.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, National Natural Science Foundation of China under contract No. 61771348, and Wuhan Morning Light Plan of Youth Science and Technology under contract No. 2017050304010302.

7. REFERENCES

[1] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *WWW*, 2010, pp. 811–820.

[2] ChaoYuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing, "Recurrent recommender networks," in *WSDM*, 2017, pp. 495–503.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[4] Paul Covington, Jay Adams, and Emre Sargin, "Deep neural networks for youtube recommendations," in *RecSys*, 2016, pp. 191–198.

[5] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li, "Temporal hierarchical attention at category-and item-level for micro-video click-through prediction," in *MM*. ACM, 2018, pp. 1146–1153.

[6] Jingwei Ma, Guang Li, Mingyang Zhong, Xin Zhao, Lei Zhu, and Xue Li, "LGA: latent genre aware micro-video recommendation on social media," *Multimedia Tools and Applications*, pp. 2991–3008, 2018.

[7] Disheng Dong, Xiaolin Zheng, Ruixun Zhang, and Yan Wang, "Recurrent collaborative filtering for unifying general and sequential recommender.," in *IJCAI*, 2018, pp. 3350–3356.

[8] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[9] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima, "Embedding-based news recommendation for millions of users," in *SIGKDD*, 2017, pp. 1933–1942.

[10] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma, "Neural attentive session-based recommendation," in *CIKM*, 2017, pp. 1419–1428.

[11] Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiusi Chen, and Jun Gao, "ATRank: An attention-based user behavior modeling framework for recommendation," in *AAAI*, 2018.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and TatSeng Chua, "Neural collaborative filtering," in *WWW*, 2017, pp. 173–182.