User-Video Co-Attention Network for Personalized Micro-video Recommendation

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

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

Hongyi Liu
Amazon Alexa, USA
liuhn@amazon.com

Xinghai Hu
Facebook Inc, USA
xinhai.me@gmail.com

ABSTRACT

With the increasing popularity of micro-video sharing where people shoot short-videos effortlessly and share their daily stories on social media platforms, the micro-video recommendation has attracted extensive research efforts to provide users with micro-videos that interest them. In this paper, a hypothesis we explore is that, not only do users have multi-modal interest, but micro-videos have multi-modal targeted audience segments. As a result, we propose a novel framework User-Video Co-Attention Network (UVCAN), which can learn multi-modal information from both user and microvideo side using attention mechanism. In addition, UVCAN reasons about the attention in a stacked attention network fashion for both user and micro-video. Extensive experiments on two datasets collected from Toffee present superior results of our proposed UVCAN over the state-of-the-art recommendation methods, which demonstrate the effectiveness of the proposed framework.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommendation, micro-video, personalization, deep learning, attention mechanism

ACM Reference Format:
Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-Video Co-Attention Network for Personalized Micro-video Recommendation. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313513

1 INTRODUCTION

In the era of mobile Internet, people are now able to shoot micro-videos (or short-videos) effortlessly and share their daily stories on social media platforms, such as Musical.ly and TikTok. With the usage of social media platforms rapidly increasing, millions of micro-videos are being generated from users of different kinds. For example, Figure 1 gives an example of micro-video on Musical.ly. The ever-increasing volume of micro-videos makes it harder for users to find the micro-video that best interests them. This has increased the demand for recommender system more than ever before.

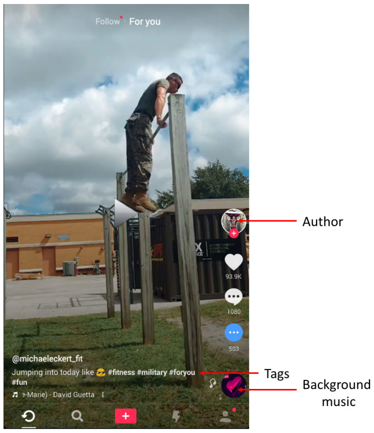

Follow
For you
Author
93.9K
1080
503
@michaeleckert_fit
Jumping into today like 😎 #fitness #military #foryou #fun
♬ e-Marie) - David Guetta
Tags
Background music

Figure 1: An example of micro-video on Musical.ly.

An offline recommender system predicts user preference and retrieves micro-video list which will be recommended to them by using users' features such as user ID or cookies, and personalizes the micro-videos for individual users [9]. However, traditional ID-based recommendation methods, such as collaborative filtering (CF) [27] and low-rank factorization [15], only consider the linear users' interaction with micro-videos without exploring non-linear and more complicated user-video interaction representation based on their features, which is important for users' hidden preferences learning. To solve this problem, recently Ma et al. [24] have proposed a genre aware neural network micro-video recommendation model on micro-video recommendation. Micro-video attention based model [5] has been explored for recommendation, where the attention mechanism learns a video representation highlighting micro-video frames relevant to user's preference. In this paper, we argue that, not only do users have multi-modal interest, but micro-videos have multi-modal targeted audience segments. In addition to visual feature, we also consider more micro-video side information that has effect on user's preference.

[*] The corresponding author.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313513

In this paper, we present a novel User-Video Co-Attention Network (UVCAN) framework for personalized micro-video recommendation which can learn multi-modal information from both user and micro-video side using attention mechanism. Unlike previous work, which connects these side information features as the user or micro-video embedding representation and only focuses on video attention, UVCAN proposes alternating co-attention to attend both user and video modalities. Through the user-video co-attention network, micro-video's and user's representations focus more on key features that demonstrate user's hidden preference. In addition, for each attention processing, we design stacked attention network to learn attention representation since only one layer of reasoning may be not enough to learn complex multi-modal features representation. Besides, we adopt a connection to associate the learned user representation, micro-video multi-modal representation, user ID embedding and micro-video ID embedding, hoping to combine collaborative filtering and attention mechanism learning for personalized micro-video recommendation.

The main contributions of this work are summarized as follows:

- A novel co-attention mechanism for personalized micro-video recommendation is proposed to jointly perform user-guided micro-video attention and video-guided user attention. Moreover, we explore this mechanism with multi-modal micro-video features and user profile.
- A stacked attention network is proposed for each attention processing to deal with complex multi-modal information, and attention learning on user attributes or micro-video multi-modal features is refined with the network step by step.
- Real-world micro-video datasets are constructed, which consist of user's attributes (user gender and location), micro-video's content information (micro-video author, background music and tags) and user-video interaction records. Extensive experiments on our collected micro-video datasets with about 336K interactions demonstrate that our proposed UVCAN outperforms current state-of-the-art methods that use hybrid approaches.

2 RELATED WORK

2.1 Video Recommendation

The existing approaches to deal with video recommendation fall into three categories: collaborative filtering [1, 11, 18], content-based filtering [9, 10] and hybrid approaches [3, 5]. Majority of the existing approaches deal with traditional videos (e.g., YouTube) while little research has been conducted for micro-videos. Because of the increasing popularity of micro-video sharing, researchers have paid more attention to micro-video data analytics. For example, Ma et al. [24] proposed a latent genre aware neural network micro-video recommendation model on micro-video recommendation. More recently, Chen et al. [7] proposed temporal attention network for user behavior modeling for micro-video recommendation. In this paper, we build new micro-video datasets involving more side information, such as user's attribute information as well as micro-video's background music and author, for studying personalized micro-video recommendation. Our proposed model using stacked co-attention deep network to sequentially model user and micro-video representations is also distinctive from the aforementioned work.

2.2 Recommendation with Neural Networks

Recently, deep neural networks are being used to extract refined latent features [14] from the user-item interaction data. The early work [26] proposes Restricted Boltzmann Machines (RBM) with an input layer and a hidden layer for collaborative filtering. AutoEncoder has been applied to recommendation tasks, e.g., AutoRec [28], DAE [22] and CDAE [33]. Cheng et al. [8] propose Wide& Deep approach for app recommendation, where the deep component is a MLP on concatenated feature embedding vectors.

Attention mechanism has been applied to deal with a variety of problem including computer vision [34], neural language processing [23] and recommender systems [2, 4, 30, 31]. For example, Chen et al. [4] introduce a attention mechanism to explore the usefulness of reviews, and propose a neural attentional regression model with review-level explanations for recommendation. Wei et al. [36] propose a user-guided attention network to attend image's multi-modal information for image representation learning. However, Chen's and Wei's work [4, 36] only utilize user-guided attention on item to generate item representation, and do not attend user side attention. These methods are fundamentally different from our proposed approach that employs user-video co-attention network to jointly describe user and micro-video attention mechanism.

3 PRELIMINARIES

In this section, we first introduce the notation and define the problem. We then introduce the embedding of user's and micro-video's information to construct the input representation.

3.1 Problem Formulation and Notations

Table 1: Definitions of notations.

Notation	description
$\mathcal{U}, \mathcal{V}$	user set, micro-video set
S^u, S^v	multi-modal information embedding of u, v
u, v	a specific user, item
$\boldsymbol{h}^u, \boldsymbol{h}^v$	identity embedding vector of u, v. $h^u, h^v \in \mathbb{R}^K$
$\boldsymbol{s}^u, \boldsymbol{s}^v$	feature embedding vector of u, v. $s^u, s^v \in \mathbb{R}^K$
K	dimensionality of the embedding vector
$\boldsymbol{Y}$	user-video interaction matrix $\boldsymbol{Y} \in \mathbb{R}^{U \times I}$
$\delta(\cdot)$	the logistic sigmoid function

We are tackling the micro-video recommendation problem by jointly considering user and micro-video attention. We consider the task of recommending micro-videos denoted as $\mathcal{V} = \{v_1, v_2, ..., v_{|\mathcal{V}|}\}$ to users denoted as $\mathcal{U} = \{u_1, u_2, ..., u_{|\mathcal{U}|}\}$. Each user u is associated with some attribute features denoted as S^u (e.g., user gender, location): $S^u = (S^u_1, S^u_2, ..., S^u_{|S^u|})$. The micro-video content information is denoted as S^v (e.g., micro-video visual content, author, background music): $S^v = (S^v_1, S^v_2, ..., S^v_{|S^v|})$. We define a user-video interaction matrix as $Y \in \mathbb{R}^{U \times I}$, where the entry y_{uv} is defined from user's implicit feedback, $y_{uv} = 1$ indicates that the user u has given the micro-video v a thumbs-up and $y_{uv} = 0$ indicates that there is no observed data about the interaction (user u, micro-video v). Our goal is to predict the user's preference for micro-videos not exposed to them before, and recommend micro-videos accordingly. Some notation frequently used throughout this paper is introduced in Table 1. Matrices are denoted with bold symbols.

3.2 Construction of Input Representation

For side information of users and micro-videos, the original representation are strings (e.g., 'gender=F' for a user, 'tag=fun' for a micro-video). We convert these sparse categorical features to low-dimensional vectors, which are referred as embedding vectors and initialized randomly. Incorporating user's identity vector $\boldsymbol{h}^u$ and side information, the embedding vectors for user u is extended as $\{\boldsymbol{h}^u, \boldsymbol{s}_1^u, \boldsymbol{s}_2^u, ..., \boldsymbol{s}_{|s^u|}^u\}$. The side information for the user is user gender and location. The side information for micro-video is tags, author id and background music id.

For micro-video visual feature, we follow the work [21] by extracting visual feature from micro-videos using pre-trained model. In particular, we extract image features at 1-frame-per-second using an Inception-v3 network [29]. We fetch the Relu activation of the last hidden layer and apply PCA (and whitening) to reduce feature dimensions to 1024 for storage and computational reasons. These frame-level features are aggregated into video-level by average pooling. For convenience of calculation, we use a single layer perceptron to convert each visual vector into a new vector that has the same dimension K as other side information feature vector. Incorporating micro-video identity vector $\boldsymbol{h}^v$ and visual feature, the embedding vectors for micro-video v is extended as $\{\boldsymbol{h}^v, \boldsymbol{s}_1^v, \boldsymbol{s}_2^v, ..., \boldsymbol{s}_{|s^v|}^v\}$, where $\boldsymbol{s}^v$ is the multi-modal information embedding for micro-video.

4 USER-VIDEO CO-ATTENTION NETWORK (UVCAN)

In this section, we first present our User-Video Co-Attention Network (UVCAN) model in detail. We then go through the learning details of UVCAN. The overall architecture of the model is shown in Figure 2.

For the user's embedding column vectors, we concatenate them to form the whole user feature representation matrix: $\boldsymbol{q}^u = concat^{(1)}($ $\boldsymbol{h}^u, \boldsymbol{s}_1^u, \boldsymbol{s}_2^u, ..., \boldsymbol{s}_{|s^u|}^u)$ where $\boldsymbol{q}^u \in \mathbb{R}^{K\times(|s^u|+1)}$. For the micro-video's embedding column vectors, we concatenate them to form the whole micro-video feature representation matrix: $\boldsymbol{q}^v = concat^{(1)}(\boldsymbol{h}^v, \boldsymbol{s}_1^v, \boldsymbol{s}_2^v,$ $..., \boldsymbol{s}_{|s^v|}^v)$ where $\boldsymbol{q}^v \in \mathbb{R}^{K\times(|s^v|+1)}$. Because not only do users have multi-modal interest, but micro-video have multi-modal targeted audience segments, we propose a user-video co-attention network that learn multi-modal information from both user and micro-video side using attention mechanism.

We first define the attention operation $\hat{x} = \mathcal{A}(\boldsymbol{X}, \boldsymbol{g})$, which takes the micro-video (or user) features $\boldsymbol{X}$ and attention guidance $\boldsymbol{g}$ derived from user (or micro-video) as inputs, and outputs the attended micro-video (or user) vector x. The operation can be denoted in the following.

$$\begin{aligned} \boldsymbol{H} &= tanh(\boldsymbol{W}_x \boldsymbol{X} + (\boldsymbol{W}_g \boldsymbol{g})\boldsymbol{I}^T), \\ \boldsymbol{a}^x &= softmax(\boldsymbol{w}_{hx}^T \boldsymbol{H} + \boldsymbol{b}), \\ \hat{\boldsymbol{x}} &= \sum a_i^x \boldsymbol{x}_i, \end{aligned} \tag{1}$$

where $\boldsymbol{I}$ is a vector with all elements to be 1. $\boldsymbol{W}_x, \boldsymbol{W}_g \in \mathbb{R}^k$ are parameters. $\boldsymbol{a}^x$ is the attention weight of feature $\boldsymbol{X}$.

4.1 User-guided Micro-video Attention

In most cases, a user is attracted by some specific micro-video features. For example, the user may pay more attention to the micro-video's visual content and author, which will play vital importance role for micro-video representation that caters user's personal preference. Hence, instead of using a global vector connecting micro-video features, we use a user-guided attention layer to filter and find features that are relevant user's personalized preference.

Before the micro-video attention, we first take user's features $\boldsymbol{q}^u$ and attention guidance $\boldsymbol{g} = \boldsymbol{0}$ to summarize the user features into a single vector $\hat{\boldsymbol{q}}_0$. Next, we take micro-video features $\boldsymbol{q}^v$ as $\boldsymbol{X}$ and the guidance $\boldsymbol{g}$ is the intermediate attended user feature $\hat{\boldsymbol{q}}_0$. Based on the attention probability of each micro-video feature, the new representation of the micro-video is constructed as weighted sum of the micro-video vector. we then use the new micro-video representation $\hat{\boldsymbol{v}}_v$ to guide the user attention.

4.2 Video-guided User Attention

Compared to models that only use user identity to guide item attention, the co-attention network model develops a more informative representation by mutually attending micro-video and user. After the phrase of user-guided micro-video attention (see Figure 2), UVCAN further incorporates user attention guided by micro-video representation since user's preference on micro-video varies with different user attributes such as gender and location. In order to obtain the user attention distribution, we take the new representation $\hat{\boldsymbol{v}}^v$ of micro-video to query the original user feature matrix $\boldsymbol{q}^u$ according to Equation 1. We then generate a new representation $\hat{\boldsymbol{v}}^u$ for user.

4.3 Stacked Attention Networks

For each attention processing, we design stacked attention network to learn attention representation which can explore some subtle relationship among query and value by iteratively querying the original feature matrix in each $\mathcal{A}$. The formula can be summarized as follows: for the m-th attention layer (where m is greater than or equal to 2), we compute the distribution of video-guided (or user-guided) by user (or micro-video) query and generate a new representation for the micro-video (or user) based on the attention probability. The new query vector is formed by adding the new feature vector to previous query vector, the equation for micro-video attention guided by user vector is as follows:

$$q_m^u = \hat{v}_m^v + q_{m-1}^u, \tag{2}$$

where q_1^u is initialized to be $\hat{q}_0$. The user attention is as same.

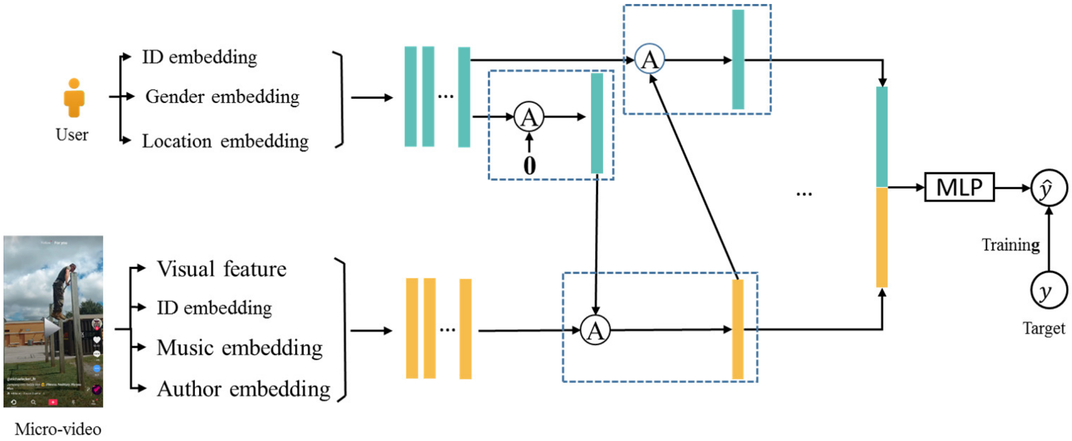


Figure 2: The illustration of Collaborative filtering stacked attentive network framework with three-step reasoning to explore user's attention to micro-video's features.

4.4 Learning for Recommendation

To test the effectiveness of learned user and micro-video representations via co-attention network on personalized micro-video recommendation, we connect these two learned multi-modal information representation vectors, user ID embedding vector and micro-video ID embedding vector to get the combination of collaborative filtering and attention mechanism. And we utilize a simple 3-layer feed-forward neural network to generate final preference rating, which dose not incur much modal complexity and ensures the capacity of nonlinear modeling for recommendation. Specifically, we define the computational formula as follows:

$$z = \begin{bmatrix} \boldsymbol{\hat{v}}^{u} \\ \boldsymbol{\hat{v}}^{v} \\ \boldsymbol{h}^{u} \\ \boldsymbol{h}^{v} \end{bmatrix}, \tag{3}$$

$$\hat{y}_{uv} = \mathrm{MLP}(z), \tag{4}$$

where the final output layer is the estimated scores $\hat{y}_{uv}$ and training is carried out by minimizing the point wise loss function between estimated score $\hat{y}_{uv}$ and real value y_{uv}.

To learn the model parameters, UVCAN is trained as a binary classification problem. Considering the one-class nature of implicit feedback, we can view $y_{ui} = 1$ as a positive label that means user u likes the micro-video v otherwise it is a negative one. The UVCAN predicts the likelihood of user u favors micro-video v. With sigmoid $\delta(\cdot)$ as the activation function for the output layer, we can constrain the output $\hat{y}_{uv}$ in $[0, 1]$. Based on the above setting, we minimized the cross-entropy loss for true labels and sampled negative instances:

$$L(y_{uv}, \hat{y}_{uv}) = \sum_{(u,v)\in Y^{+}\cup Y^{-}} -y_{uv}\log(p) - (1-y_{uv})\log(1-p), \tag{5}$$

where $p = \delta(\hat{y}_{uv}) = 1/(1 + exp(-\hat{y}_{uv}))$ represents the sigmoid function on the output $\hat{y}_{uv}$. Y^{+} and Y^{-} denote the observed user-video pairs and unobserved user-video pairs, respectively. We uniformly sample negative instances Y^{-} from unobserved interaction according to the fixed sampling ratio 4:1 to the number of observed interactions in each iteration. An Adaptive Moment Estimation (Adam) [19] algorithm is adopted to optimize the loss function.

5 EXPERIMENTS

In this section, we conduct extensive experiments on the collected micro-video datasets from Toffee to compare UVCAN against other state-of-the-art recommendation methods. We first briefly depict the collected two real-world micro-video datasets, followed by experimental setting and evaluation scheme. Then we introduce the baseline algorithms and parameter settings. Finally, we present our experimental results and discussions, with comparison to different kinds of baseline methods.

5.1 Dataset

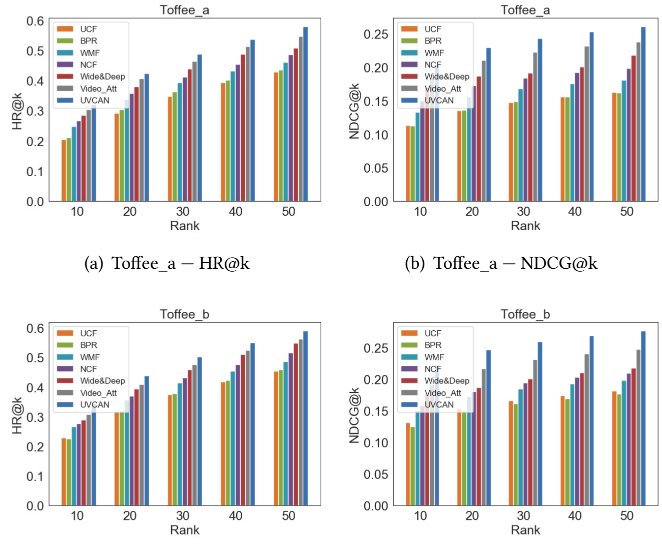
Table 2: Basic statistics of datasets.

Dataset	#User	#Micro-video	#Interaction	Span
Toffee_a	780	73,015	233,144	2017.01-2017.06
Toffee_b	3,231	50,574	336,460	2017.07-2018.06

We construct our datasets using micro-videos collected from Toffee, one of China's popular micro-video social media platform. As far as we know, there is no publicly available micro-video social media dataset that contains users' metadata, content information of micro-videos, and the user-to-video connections simultaneously. In addition to establish connections (make new friends) between users, Toffee allows users to publish and share videos to others, and send thumbs-up to different videos. Toffee is a micro-video social media app which starts run from Jan 2017. To study the effect of different phrase of development on recommendation, we crawled micro-videos of Jan 2017 to Jun 2017 as the developing phrase and Jul 2017 to Jun 2018 as the plateau.

We constructed the two datasets from the Toffee in a user-centric method, which is also adopted by [6]. We first crawled one week of public timeline micro-videos and random sampled active users.

We defined active users who have at least interacted (send a thumb-up to the video to express liking) with 5 micro-videos. Then we expanded the seed users by crawling their followees, as the followees of active users were more likely to be active users relative to their followers. We performed three layers of crawling and further crawled users' ID and registering information, the ID list of their preferred micro-videos and the content information of these micro-video such as video file, tag, author ID and music ID. To test a scenario where user profile is known, we filtered out users without location or gender information. With this crawling method, we got the two datasets for different developing phrase: Toffee_a and Toffee_b. The detailed statistics of these datasets after preprocessing are shown in Table 2. The sparsity ($\#interaction/(\#user \times \#video)$) of the two datasets is 0.4094% for Toffee_a and 0.2059% for Toffee_b.



(a) Toffee_a — HR@k

(b) Toffee_a — NDCG@k

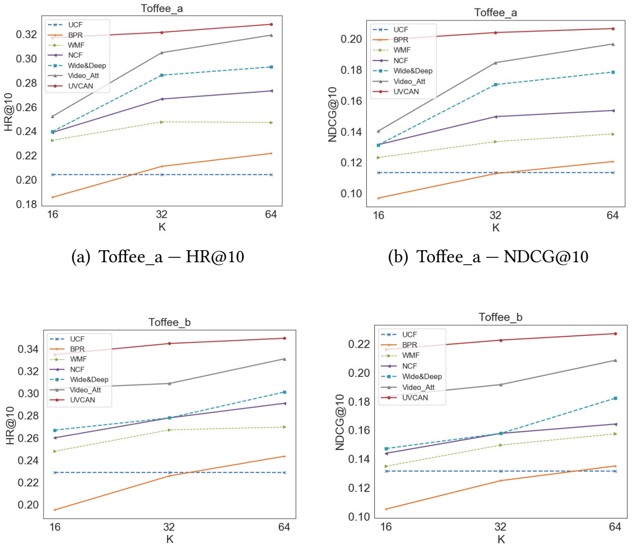
(c) Toffee_b — HR@k

(d) Toffee_b — NDCG@k

Figure 3: Performance of top-k recommendation where k range from 10 to 50 on Toffee_a.

5.2 Evaluation Scheme

For each dataset, we randomly hold 20% of micro-videos associated with each user to form the test set, and put the other micro-videos in the training set. The splitting methods is widely used in previous work on recommendation system [16, 32, 35]. In term of performance evaluation, for each (u, v) pair in the test set, we random select 1000 additional micro-video and predict scores by u for v and the other 1000 micro-videos. The evaluation criteria is to measure how well the method ranks the correct pair (u, v) against the other random micro-videos. This evaluation strategy is commonly used by [12, 13, 20]. We evaluate our method and other compared ones on micro-video recommendation by using two popular metrics, i.e., Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG), which regard liking as positive labels.



(a) Toffee_a — HR@10

(b) Toffee_a — NDCG@10

(c) Toffee_b — HR@10

(d) Toffee_b — NDCG@10

Figure 4: Performance of HR@10 and NDCG@10 w.r.t dimensionality of predictive factors on two datasets.

5.3 Baselines

We compare our proposed UVCAN with a series of state-of-the-art baseline algorithms as follows: **UCF** [37], **BPR** [25], **WMF** [17], **NCF** [14], **Wide&Deep** [8] and **Video-Att**. **Video-Att** is a variant of our proposed model, which only use user-guided micro-video attention. Similar work is done by [5], which proposes a linear model considering user-guided attention to item while **Video-Att** extends the implement to non-linear model.

5.4 Parameter Settings

We use a Gaussian distribution to randomly initialized the model parameters with a mean of 0 and standard deviation of 0.1 for our UVCAN model, and optimize the model through mini-batch Adaptive Moment Estimation (Adam). We tested the learning rate of [0.0001, 0.001, 0.01, 0.1], latent feature dimension of [16, 32, 64] and the regularizer of [0, 0.001, 0.01, 0.1]. The grid search method is used to determine hyper-parameters. If not specified, we employed three hidden layers for multi-Layer perceptron and show the results of $K = 32$ for all methods because of the consistent across the dimension of latent vector. Without special mention, we employ two attention layers for UVCAN.

5.5 Results and Analysis

5.5.1 Model Comparison. The performance of all competitors with HR@10 and NDCG@10 on micro-video recommendation is shown in Table 3. Figure 3 demonstrates the performance of Top-k recommendation list with ranking position starting from 10 to 50 on Toffee_a and Toffee_b dataset. From the table and figure, we have the following observations:

Table 3: Recommendation performance in terms of the HR@10 and NDCG@10.

Dataset		(a)	(b)	(c)	(d)	(e)	(f)	(g)	Improvement		
		UCF	BPR	WMF	NCF	Wide&Deep	Video-Att	UVCAN	(g) vs. (c)	(g) vs. (e)	(g) vs. Best
Toffee_a	HR10	0.2043	0.2110	0.2478	0.2666	0.2863	0.3049	**0.3217**	29.81%	12.36%	5.50%
	NDCG10	0.1135	0.1128	0.1335	0.1497	0.1705	0.1847	**0.2042**	52.92%	19.74%	10.53%
Toffee_b	HR10	0.2290	0.2259	0.2672	0.2780	0.2891	0.3089	**0.3448**	29.06%	19.28%	11.64%
	NDCG10	0.1317	0.1250	0.1497	0.1578	0.1765	0.1917	**0.2226**	48.72%	26.13%	16.13%

(a) On average, our proposed method outperforms the second-best method by 8.57% in terms of HR@10 and 13.33% in terms of NDCG@10 across the two datasets. We speculate that learning users' hidden preferences through combining micro-video and user mutual attention mechanism improves the performance of recommendation, which can be also seen from the superior performance of UVCAN than that of Video_Att.

(b) Using attention networks to explore a more powerful latent representation for user and micro-video improves the performance of recommendation. UVCAN and Video_Att outperform the model that concatenates identity embedding and side feature embedding of user and micro-video and feeds them directly in deep neural networks (namely, NCF and Wide&Deep).

(c) As can be seen from Figure 3, UVCAN shows consistent improvement compared with other methods in different ranking positions. While the Toffee_b dataset is more sparse than Toffee_a, the performance of all competitor on Toffee_b is much better. The possible reason may be that in Toffee_b dataset, which is from the developing plateau of the social media platform, the connection between users is constructed more densely, which also explain why the performance of UCF on Toffee_b is better than that of BPR.

5.5.2 Effects of the number of predictive factors. Figure 4 shows the performance of HR@10 and NDCG@10 with respect to different size of predictive factors on two datasets. When other factors remain the same, increasing the number of predictive factors introduces more information. As we can observe from Figure 4, compared to other competitor methods, the performance improvement of UVCAN also increases substantially. The possible reason is that the combination of collaborative filtering and attention mechanism requires larger embedding vector to incorporate the abundant information. For Toffee_b, even with a small predictive factor of 16, UVCAN performs substantially perform better than WMF and BPR with a large factor of 64. This shows the high effectiveness of UVCAN by learning users' hidden preferences.

5.5.3 Effects of the number of attention layers. We experiment with different number of attention layers to investigate the effect on the recommendation performance of UVCAN on Toffee_a and Toffee_b datasets, which is shown in Tables 4 and 5. We only show the results of NDCG, and the results of HR admit the same trend thus they are omitted. AL-2 indicates the micro-video attention and user attention both have two attention layers with other factors remaining the same. As can be seen, the performance improves at the beginning which also indicates the effectiveness of stacked attention layers. We owe this improvement to the multi-step reasoning of user-guided attention on micro-video's side information and video-guided attention on user's attribute information. When using more attention layers than two or three, the performance does not further improve significantly.

Table 4: NDCG@10 of UVCAN with different attention layers on Toffee_a.

factors	AL-1	AL-2	AL-3	AL-4
16	0.1947	**0.1993**	0.1948	0.1920
32	0.2022	**0.2042**	0.1927	0.1982
64	0.1947	0.1968	**0.2009**	0.2008

Table 5: NDCG@10 of UVCAN with different attention layers on Toffee_b.

factors	AL-1	AL-2	AL-3	AL-4
16	0.2045	**0.2161**	0.2123	0.2124
32	0.2161	**0.2226**	0.2116	0.2094
64	0.2105	0.2171	**0.2219**	0.2159

6 CONCLUSION

In this paper, we propose a User-Video Co-Attention Network (UVCAN) framework for personalized micro-video recommendation, based on the hypothesis that not only do users have multi-modal interest but micro-videos have multi-modal targeted audience segments. UVCAN achieves superior performance over state-of-the-art methosd by following advantages: 1) It uses embedding-based method to represent side information features and users' history interaction behavior with micro-videos; 2) It explores users' hidden preferences on side information of micro-videos, which views user's features as an input query and learn micro-video attention through multi-step reasoning with stacked attention network. 3) In addition, it explores user's diversity preference from different attribute, which views the learned micro-video representation as query and learn user attention through multi-step reasoning with stacked attention network. We conduct extensive experiments on two collected real-world micro-video datasets to compare UVCAN with state-of-the-art methods and demonstrate its effectiveness.

ACKNOWLEDGEMENTS

This work was supported in part by National Key R&D Program of China under contract No. 2017YFB1002202, and National Natural Science Foundation of China under contract No. 61771348.

REFERENCES

[1] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web(WWW'08)*. ACM, 895–904.

[2] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. 2018. Attentive Group Recommendation. In *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. ACM.

[3] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. 2012. Personalized video recommendation through tripartite graph propagation. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. ACM, 1133–1136.

[4] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 27th International Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee.

[5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and TatSeng Chua. 2017. Attentive collaborative filtering: multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. ACM, 335–344.

[6] Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *Proceedings of the 24nd ACM International Conference on Multimedia (MM'16)*. ACM, 1018–1027.

[7] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal Hierarchical Attention at Category-and Item-Level for Micro-Video Click-Through Prediction. In *Proceedings of the 26nd ACM International Conference on Multimedia (MM'18)*. ACM, 1146–1153.

[8] HengTze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*. ACM, 7–10.

[9] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*. ACM, 191–198.

[10] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM'14)*. ACM, 597–606.

[11] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*. ACM, 293–296.

[12] Chao Du, Chongxuan Li, Yin Zheng, Jun Zhu, and Bo Zhang. 2018. Collaborative Filtering with User-Item Co-Autoregressive Models. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI'18)*. AAAI Press.

[13] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web (WWW'13)*. International World Wide Web Conferences Steering Committee, 278–288.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and TatSeng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web(WWW'17)*. International World Wide Web Conferences Steering Committee, 173–182.

[15] Xiangnan He, Hanwang Zhang, MinYen Kan, and TatSeng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, 549–558.

[16] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. International World Wide Web Conferences Steering Committee, 193–201.

[17] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 15th IEEE International Conference on Data Mining (ICDM'08)*. IEEE, 263–272.

[18] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. 2016. Real-time video recommendation exploration. In *Proceedings of the 2016 International ACM SIGMOD Conference on Management of Data (SIGMOD'16)*. ACM, 35–46.

[19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR'2015)*.

[20] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'08)*. ACM, 426–434.

[21] Joonseok Lee, Sami AbuElHaija, Balakrishnan Varadarajan, and Apostol Paul Natsev. 2018. Collaborative Deep Metric Learning for Video Understanding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'18)*. ACM, 481–490.

[22] Sheng Li, Jaya Kawale, and Yun Fu. 2015. Deep collaborative filtering via marginalized denoising auto-encoder. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*. ACM, 811–820.

[23] Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content Attention Model for Aspect Based Sentiment Analysis. In *Proceedings of the 27th World Wide Web Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee, 1023–1032.

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

[25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*. AUAI Press, 452–461.

[26] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. 2007. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*. 791–798.

[27] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Darius Braziunas. 2016. On the Effectiveness of Linear Models for One-Class Collaborative Filtering. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 229–235.

[28] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th International Conference on World Wide Web(WWW'15)*. International World Wide Web Conferences Steering Committee, 111–112.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1–9.

[30] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Latent relational metric learning via memory-based attention for collaborative ranking. In *Proceedings of the 27th World Wide Web Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee, 729–739.

[31] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-Pointer Co-Attention Networks for Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'18)*. ACM.

[32] Keqiang Wang, Yuanyuan Jin, Haofen Wang, Hongwei Peng, and Xiaoling Wang. 2018. Personalized Time-Aware Tag Recommendation.. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI'18)*. AAAI Press.

[33] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*. ACM, 153–162.

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'2015)*. 2048–2057.

[35] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee, 649–658.

[36] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 27th World Wide Web Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee, 1277–1286.

[37] ZhiDan Zhao and MingSheng Shang. 2010. User-based collaborative-filtering recommendation algorithms on hadoop. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'10)*. ACM, 478–481.