Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction*

Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, Yan Li
CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System
University of Science and Technology of China
Hefei, China

ABSTRACT

Micro-video sharing gains great popularity in recent years, which calls for effective recommendation algorithm to help user find their interested micro-videos. Compared with traditional online (e.g. YouTube) videos, micro-videos contributed by grass-root users and taken by smartphones are much shorter (tens of seconds) and more short of tags or descriptive text, making the recommendation of micro-videos a challenging task. In this paper, we investigate how to model user's historical behaviors so as to predict the user's click-through of micro-videos. Inspired by the recent deep network-based methods, we propose a Temporal Hierarchical Attention at Category- and Item-Level (THACIL) network for user behavior modeling. First, we use temporal windows to capture the short-term dynamics of user interests; Second, we leverage a category-level attention mechanism to characterize user's diverse interests, as well as an item-level attention mechanism for fine-grained profiling of user interests; Third, we adopt forward multi-head self-attention to capture the long-term correlation within user behaviors. Our proposed THACIL network was tested on MicroVideo-1.7M, a new dataset of 1.7 million micro-videos, coming from real data of a micro-video sharing service in China. Experimental results demonstrate the effectiveness of the proposed method in comparison with the state-of-the-art solutions.

CCS CONCEPTS

• **Information systems → Personalization**;

KEYWORDS

attention mechanism; click-through prediction; micro-video; recommendation algorithm; user modeling.

ACM Reference Format:
Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, Yan Li. 2018. Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction. In *MM '18: 2018 ACM Multimedia Conference, Oct. 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240617

1 INTRODUCTION

Micro-video sharing has become a very popular service recently. People use their smartphones to capture short (usually tens of seconds) videos, and upload them to an online sharing platform. Browsing and interacting with micro-videos in the online platform also attract many silent users that contribute less. In a micro-video sharing service in China, there are more than 60 million daily active users, and 10 millions of micro-videos are uploaded every single day. Because of the clear information overload, it is an urgent need to develop recommendation system for micro-videos in order to help users find their favorites and to enhance user experience.

Traditionally, personalized recommendation algorithms can be categorized into content-based filtering [7, 10, 20, 22, 36, 37], collaborative filtering [2, 17], and hybrid approaches [3, 4, 11, 31, 34]. Content-based filtering (CBF) requires evaluating the similarity /correlation between items according to their content (including auxiliary information such as metadata or tags), and recommends the items similar/correlated to user's historically accessed items. Collaborative filtering (CF) leverages crowd wisdom and learns from multiple users' interaction data, where interaction can be explicit (like user gives rating on item) or implicit (like user clicks item). Both CBF and CF are for personalized recommendation, so they utilize a user's historical behaviors to recommend to that user. In addition, CBF depends on content analysis which can be computationally expensive for unstructured data (audio, image, video, etc.). CF gets rid of content analysis, and instead depends on multiple users' historical behaviors. CF cannot recommend *new* item–that has not been accessed by any user. Hybrid approaches are proposed to combine the advantages of CBF and CF and other (non-personalized) algorithms, and are widely adopted in practical recommendation systems. In addition, classic recommendation algorithms usually assume that user's interests are static, whereas more and more recent works take into account the temporal dynamics of user interests, and report improved adaptivity in making recommendations.

Specific to video recommendation, recent years have witnessed much progress [6, 7, 9, 12, 20], while this task is still very difficult. Compared with the items having structured data/metadata, like books, movies, music, merchandise, etc., videos usually lack metadata and they themselves have no immediate semantic representation, which causes the problem of "semantic gap" that is hard to overcome[1]. Furthermore, micro-videos are different from traditional online (e.g. YouTube) videos. Due to the very low entry bar of content generation, micro-videos are more contributed by

[*] Corresponding author: Dong Liu (dongeliu@ustc.edu.cn).

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240617

[1] Here, movies refer to professionally produced films that usually have structured metadata (genre, director, actor/actress, etc.), videos refer to online videos that usually do not have such metadata.

grass-root users. Thereby, and also due to the mobile app nature, user-generated micro-videos are much shorter, and more short of tags or descriptive text, than traditional online videos. Thus, it is a more challenging task to recommend micro-videos.

In this paper, we propose a personalized recommendation algorithm for micro-videos. Our specific objective is to predict whether or not a given user will click a given micro-video. Once having the predicted click-through likelihood, it is straightforward to recommend those micro-videos having higher probabilities. Considering that micro-videos are extremely short of auxiliary information (tags, description, and so on), we restrict ourselves to use *merely* the visual information of video frames. More specifically, we are restricted to use only the cover picture of micro-video[2]. In addition, since micro-videos are generated at a very fast speed (e.g. 10 million every day), new micro-videos can hardly have any interaction data before they are recommended, which virtually *prohibits* the usage of CF kind of algorithms. Thus, we consider a CBF like algorithm, and we wish to improve the user behavior modeling part of our recommendation algorithm.

In the recent three years, modeling user behaviors with deep networks, especially recurrent neural network (RNN), has been an emerging topic [12, 14, 26, 29, 32]. RNN was known to be powerful in sequential modeling than the other methods like Markov chain based [24], and was widely applied into natural language processing [21] and computer vision [19]. However, the capacity of RNN is restricted when dealing with very long sequences, due to both training difficulty and high complexity of inference. To solve this problem, Zhou *et al.* [35] proposed a self-attention mechanism to exploit the long-term correlation between user behaviors. The self-attention can also be regarded as a manner to identify non-local similarity, which was also investigated in the task of video classification [28].

Inspired by the success of the aforementioned works, we propose a deep network-based method for micro-video click-through prediction. The focus of our method is a computationally efficient manner to model a user's historical behaviors as a temporal sequence.

- First, we use *temporal* windows to capture the short-term dynamics of user interests. By splitting user's behaviors into multiple segments, not only the computational load is less but also the short-term attention mechanism works better.
- Second, within each window, we design a *category- and item-level attention* mechanism to characterize user interests. The category-level attention is to describe user's diverse interests, and the item-level attention is to profile fine-grained user interests.
- Third, we propose a forward multi-head *self-attention* mechanism to identify and integrate the long-term correlation between the previously split windows. Thus, both short-term and long-term properties of user behaviors are modeled by the *hierarchical attention* mechanism.

Our network is named Temporal Hierarchical Attention at Category- and Item-Level (THACIL) to highlight its technical contributions. We test the THACIL network on MicroVideo-1.7M, a new dataset coming from real data and consisting of 1.7 million micro-videos. Experimental results demonstrate the effectiveness of the proposed method in comparison with the state-of-the-art solutions.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 details our proposed THACIL network. Experimental settings and results are reported in Section 4, followed by conclusions in Section 5.

2 RELATED WORK

2.1 Video Recommendation

Existing video recommendation methods usually fall into one of the three categories: collaborative filtering [2, 17], content-based filtering [7, 10, 20, 22, 36, 37], and hybrid approaches [3, 4, 11, 31, 34]. Regarding CF methods, for example, Baluja *et al.* [2] proposed to propagate preference information through a variety of graphs to provide personalized video recommendation. Huang *et al.* [17] developed a scalable online CF algorithm based on matrix factorization, with an adjustable updating strategy to account for implicit feedback of different user actions. CF methods suffer from the cold-start problem, i.e. when a new video is added into the library, recommender system needs to bootstrap with no or little user action on it. One promising approach for solving cold-start problem is to exploit video content, by analyzing the audio, visual cues, subtitle, tags, and so on. CBF methods are based on the analysis of video content to evaluate the similarity/correlation between videos, and to recommend new videos that are similar/correlated to user's historically accessed videos. For example, Mei *et al.* [20] presented a contextual video recommendation system based on multimodal content relevance and user feedback. Deldjoo *et al.* [10] proposed a content-based recommender system by extracting a set of representative stylistic features, including lighting, color, and motion, from videos. However, CBF methods usually suffer from the difficulty and computational complexity of video content analysis. Some hybrid approaches combine CF and CBF into a single framework. Zhao *et al.* [34] proposed a multi-task rank aggregation approach, which defined video recommendation as a ranking problem, and generated then fused multiple ranking lists by exploring different information sources. In [11], user's self-expression as found in user profile and perception of visual saliency in videos were both exploited to enhance video recommendation.

Most of the existing methods deal with traditional online (e.g. YouTube) videos, but little research has thus far been conducted for micro-videos. Due to the increasing popularity of micro-video sharing, researchers have noticed the value of micro-video-related data analytics. For example, Chen *et al.* [4] proposed a transductive model to find the optimal latent common space, unifying and preserving information from different modalities, for predicting micro-video popularity. More recently, Huang *et al.* [16] proposed a personalized micro-video recommendation method that models hierarchical user interests based on multimodal features. In this paper, we build a new and the largest known dataset for studying personalized micro-video click-through prediction. Our proposed approach using deep network and attention mechanism is also distinctive from the aforementioned works.

[2] Since micro-video is short, it usually contains no scene change, so the cover picture has most of the visual information in micro-video. In addition, in a micro-video mobile app, cover picture is the most important cue in the user interface when a user decides to click a micro-video.

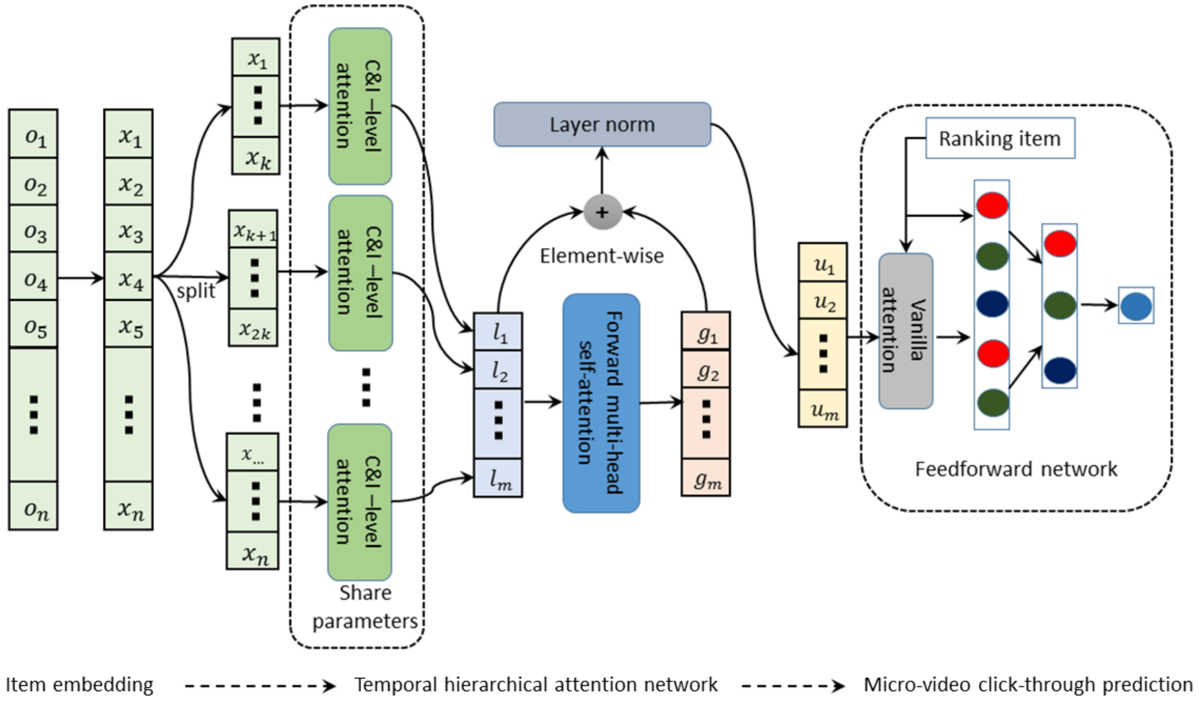


Figure 1: The proposed Temporal Hierarchical Attention at Category- and Item-Level (THACIL) network for micro-video click-through prediction. "C&I-level" stands for category- and item-level. The forward multi-head self-attention module is further shown in Fig. 2.

2.2 RNN- and Attention-Based Recommendation

Along with the great success of deep learning, deep network-based methods, especially recurrent neural network (RNN) based, have been proposed for personalized recommendation. For example, RNN-based methods were adopted for next basket recommendation [32], shopping items recommendation [14], news recommendation [26], movie recommendation [29], and so on. For video recommendation, Gao *et al.* [12] proposed a dynamic RNN to model user's dynamic interests over time in a unified framework; Hidasi *et al.* [14, 15] applied RNNs into session-based video recommender systems. However, the capacity of RNNs is claimed to be restricted when dealing with very long sequences, due to both training difficulty and high complexity of inference. Self-attention mechanism was proposed to replace RNN, and was shown to be faster and more powerful to characterize long-term dependency.

Attention mechanism has been applied to solve a variety of problems including computer vision [30], neural language processing [1, 8, 27], and recommender systems [5, 35]. For example, Bahdanau *et al.* [1] firstly introduced an attention mechanism to provide more accurate alignment for each position in the encoder-decoder framework for machine translation task. Shen *et al.* [25] proposed a directional self-attention mechanism, which helped achieve the state-of-the-art performance in language understanding. Witnessing the success of multi-head self-attention in machine translation task [27], Zhou *et al.* [35] exploited self-attention mechanism to model user behaviors, each of which was considered influenced by other behaviors, and reported significant improvement than RNN-based methods. In addition, Chen *et al.* [5] introduced multi-level attention, i.e. item- and component-level attention, in the CF framework for image/video recommendation.

In this paper, we propose a temporal hierarchical attention mechanism for micro-video recommendation. We split user's historical behaviors into multiple segments, apply a short-term attention model on each segment, and further apply a long-term attention model between the multiple segments. For the short-term attention, we adopt multi-level, i.e. category- and item-level attention. For the long-term attention, we adopt forward multi-head self-attention. To the best of our knowledge, we are the first to explore a unified attention mechanism for both short- and long-term dependency in modeling user behaviors for personalization purpose.

3 METHODS

For micro-video recommendation, as there is a severe cold-start problem, we retreat to a CBF like algorithm. Specifically, we assume that user's preferences can be discovered from the micro-videos he/she has accessed, and we want to evaluate the similarity/correlation between a new micro-video and user's accessed micro-videos, so as to predict the click-through probability. Here, different from traditional CBF methods that ignore the temporal/sequential information, we regard one user's accessed micro-videos as a sequence in the temporal order. In summary, we are given a sequence of items (to represent a user) as well as a new item, and we are to predict the user-item interaction probability.

Micro-video has multiple attributes that belong to multiple modalities, including audio, visual cues, tags, and so on. In this work we consider only two kinds of information–features extracted from *cover picture*, and *category*. Extensions to other modalities are straightforward, and will be considered in our future work.

Our proposed THACIL network is illustrated in Figure 1. It consists of three parts. The first part converts a sequence of items into their representations. The second part works on the sequence with temporal hierarchical attention mechanism. And the third part accepts a new item as input to predict its corresponding click-through probability. The three parts are discussed in the following three subsections, respectively.

3.1 Item Embedding

We seek a uniform manner to convert each item into its representation, i.e. embed an item into a d-dimensional space. Thus, a sequence of videos, denoted by $\{o_1, o_2, \ldots, o_n\}$ in Fig. 1, are converted to a sequence of d-dimensional vectors, denoted by $\{x_1, x_2, \ldots, x_n\}$.

As mentioned above, for each micro-video o_i, we have features extracted from its cover picture, i.e. o_i^f, and its category o_i^c. In this paper, o_i^f is produced by using the Inception-v3 model pretrained on ImageNet on cover picture[3]. As o_i^f is high-dimensional, we perform a linear embedding for it,

$$f_i = E_f o_i^f \tag{1}$$

where $f_i \in \mathbf{R}^{d_f}$ is an embedded vector at the dimension of d_f.

In addition, each micro-video belongs to one and only one category[4]. The category can be denoted as a one-hot vector o_i^c for o_i. We also train a linear embedding for it,

$$c_i = E_c o_i^c \tag{2}$$

where $c_i \in \mathbf{R}^{d_c}$ is an embedded vector at the dimension of d_c.

At last, we concatenate the embedded vectors of cover picture and category, i.e. $x_i = [f_i; c_i]$. Accordingly, $x_i \in \mathbf{R}^d$ where $d = d_f + d_c$. Please note that the embedding matrices E_f and E_c are trained out.

3.2 Temporal Hierarchical Attention

One user's historically accessed items can be many. If we use RNN to analyze the sequence $\{x_1, x_2, \ldots, x_n\}$, we may fail to capture both short-term and long-term dependency in it. Thus, we intentionally perform a multi-scale analysis using the proposed temporal hierarchical attention mechanism that will be detailed in this subsection.

First, we split the sequence into m blocks where each block contains k items ($m \times k = n$). In each block, we apply the category- and item-level attention to achieve $l_i \in \mathbf{R}^d$. Here l_i encodes the *local* information within the i-th block ($i = 1, 2, \ldots, m$). We then apply the forward multi-head self-attention to exploit the correlation between the blocks, and achieve $g_i \in \mathbf{R}^d$, where g_i encodes kinds of *global* information from the 1-st to the i-th block (due to the forward nature, as detailed below). Local information and global information are combined and normalized to achieve $\{u_1, u_2, \ldots, u_m\}$, which can be perceived as the user profile built from a sequence of items.

Category- and item-level attention. In this subsection, we use $x_{ij}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$ to denote the j-th item in the i-th temporal block. Our category-level attention score is calculated as

$$\alpha_c(c_{ij}) = W_c \sigma(W_1 c_{ij} + W_2 f_{ij} + b_1) + b_2 \tag{3}$$

where the weight matrices W_c, W_1, W_2 are of shapes $\mathbf{R}^{d_c \times d_c}, \mathbf{R}^{d_c \times d_c}, \mathbf{R}^{d_c \times d_f}$, respectively, bias vectors b_1, b_2 are both of dimension d_c, $\sigma(\cdot)$ is the element-wise activation function. The attention scores are then normalized using the *softmax* function,

$$\tilde{\alpha}_c(c_{ij}) = \frac{\exp(\alpha_c(c_{ij}))}{\sum_{j=1}^{k} \exp(\alpha_c(c_{ij}))} \tag{4}$$

And the normalized scores are used to fuse the category information of items within a block, i.e. $l_i^c = \sum_{j=1}^{k} \tilde{\alpha}_c(c_{ij}) \odot c_{ij}$, where $\odot$ represents element-wise product. Similarly, our item-level attention score is calculated as

$$\alpha_f(f_{ij}) = W_f \sigma(W_1' c_{ij} + W_2' f_{ij} + b_1') + b_2' \tag{5}$$

where the weight matrices W_f, W_1', W_2' are of shapes $\mathbf{R}^{d_f \times d_f}, \mathbf{R}^{d_f \times d_c}, \mathbf{R}^{d_f \times d_f}$, respectively, bias vectors b_1', b_2' are both of dimension d_f, $\sigma(\cdot)$ is the element-wise activation function. The softmax normalized scores are used to fuse the visual information of items within a block, i.e. $l_i^f = \sum_{j=1}^{k} \tilde{\alpha}_f(f_{ij}) \odot f_{ij}$. At last, l_i^c and l_i^f are concatenated to form $l_i \in \mathbf{R}^d$.

Forward multi-head self-attention. We regard self-attention [27] as a mean to exploit the correlation between different temporal blocks. Moreover, we consider forward multi-head self-attention to restrict the information flow from old to new. As illustrated in Fig. 2, we first duplicate the local information $\{l_i, i = 1, 2, \ldots, m\}$ three times to be $\{q_i\}$, $\{k_i\}$, and $\{v_i\}$. q_i and k_i are used to calculate attention scores that will be applied onto v_i. Specifically, we linearly project q_i (resp. k_i and v_i) h times, each time with a different projection matrix to the dimension of d_q (resp. d_k and d_v). We set $d_q = d_k = d_v = d/h$. At the p-th time, we produce a d_v-dimensional attention score between the i- and j-th blocks,

$$H_{i,j}^{(p)} = W_p \sigma(W_p^q q_i + W_p^k k_j + b_p) \tag{6}$$

where $p = 1, 2, \ldots, h$, W_p^q and W_p^k are the projection matrices for q_i's and k_i's, and W_p is a weight matrix of shape $\mathbf{R}^{d_v \times d_q}$, b_p is a bias vector of dimension d_q. This attention score is then added by a directional mask $\mathbf{M}$, whose elements are defined as

$$M_{i,j}^{(p)} = \begin{cases} 0, & \text{if } i > j \\ -\infty, & \text{otherwise} \end{cases} \tag{7}$$

The mask together with the following softmax function will make the contribution from j to i be 0 if $j \geq i$, thus enforcing *forward* attention that only enables contribution from j to i once $j < i$. The

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

[4] Categorical information is manually defined and provided in our dataset, as will be detailed later.

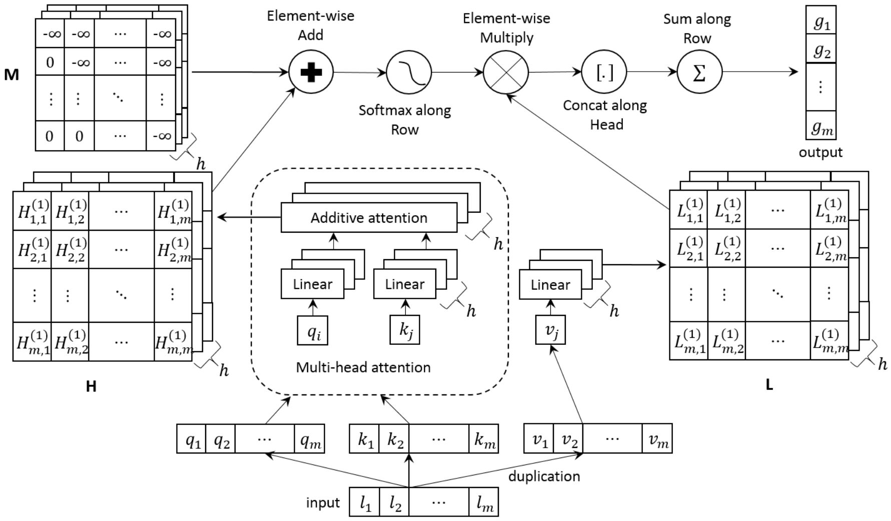


Figure 2: The proposed forward multi-head self-attention mechanism.

softmax normalized attention scores are used on $\mathbf{L}$, whose elements are defined as

$$L^{(p)}_{i,j} = W^v_p v_j \tag{8}$$

where W^v_p is the projection matrix for v_i's. After attention weighting, the tensor $\mathbf{L}$ is reorganized by concatenating along the p index (resulting in a tensor of shape $m \times m \times d$ because $d = d_v \times h$), and then by summing over the i index, resulting in $\{g_1, g_2, \ldots, g_m\}$.

3.3 Micro-Video Click-Through Prediction

From the above analyses, we have reached user profile as a sequence $\{u_1, u_2, \ldots, u_m\}$, where each $u_i \in \mathbf{R}^d$. In this section, we predict the click-through probability using the sequence and a new item. The new item is also embedded into a d-dimensional vector, using the trained embedding matrices as mentioned in Section 3.1. Denote the new item's embedded vector by x.

Given x, we first calculate an attention score on the user profile, i.e.

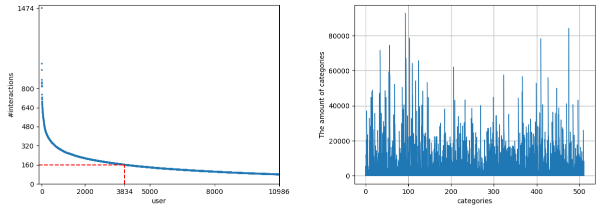
$$\alpha_u(u_i) = W_u \sigma(W_3 u_i + W_4 x + b_3) \tag{9}$$

where the weight matrices W_u, W_3, W_4 are all of shape $\mathbf{R}^{d \times d}$, bias vector b_3 is of dimension d. This attention score is softmax normalized and then used on u_i, so as to achieve $u = \sum_{i=1}^m \tilde{\alpha}_u(u_i) \odot u_i$. We then concatenate u and x, and pass them to a feedforward network of two layers, to predict the click-through probability.

The entire network shown in Fig. 1 can be trained end-to-end. During training, it is natural to use the sigmoid cross-entropy loss function:

$$L(u, x) = y \log \sigma(f(u, x)) + (1 - y) \log(1 - \sigma(f(u, x))) \tag{10}$$

where $y \in \{0, 1\}$ is the ground-truth that indicates whether the user clicks the micro-video or not, and f represents the feedforward network.



(a) Distribution of user clicked micro-video counts. **(b) Distribution of category-wise micro-video counts.**

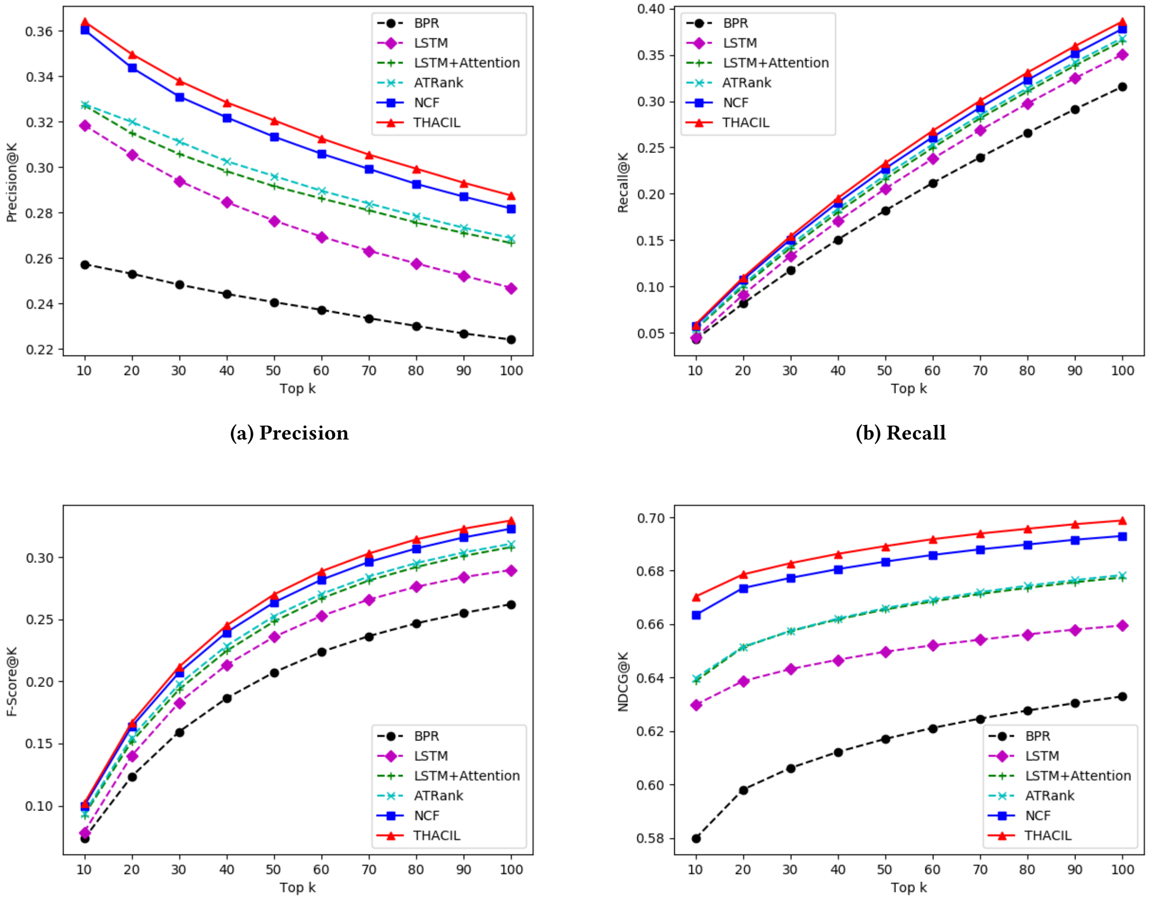
Figure 3: Distributions of MicroVideo-1.7M training set.

4 EXPERIMENTS

4.1 Dataset

We build a new dataset for our study, using real data from a famous micro-video sharing service in China[5]. Our built dataset, named MicroVideo-1.7M, has 12,737,619 interactions that 10,986 users have made on 1,704,880 micro-videos. For each micro-video, features have been extracted from its cover picture using the Inception-v3 model, as mentioned in Section 3.1. In addition, there is a manually designed categorization for micro-videos with 512 categories, where each micro-video belongs to one and only one category. Each interaction has its associated user ID and micro-video ID, and a timestamp. Note that the timestamps have been processed so that the absolute time is unknown, but the sequential order is preserved. There are not only "positive" interactions, i.e. user clicks micro-video, but also "negative" interactions, i.e. user has observed cover picture (in the user interface in the mobile app) but does not click. Thus, the dataset is very suitable for studying click-through prediction. The dataset and our source code can be accessed at https://github.com/Ocxs/THACIL.

[5] The service provider helps us build the dataset, but requires anonymity.



(a) Precision

(b) Recall

(c) F-score

(d) NDCG

Figure 4: Quantitative results of different methods on the MicroVideo-1.7M test set.

In practice, new micro-videos are generated at a very high speed, so we intentionally study the item cold-start scenario. That says, we divide the micro-videos into two disjoint sets, and divide the interactions according to the micro-videos into two sets, one for training and the other for test. Statistics of the divided dataset are shown in Table 1, note that the users are exactly the same in the training and test sets.

To better profile the dataset, we draw two distributions in Fig. 3. The distribution of user clicked micro-video counts shows long tail, i.e. a few users are very active but many are not that active. Note that the average number of clicked micro-videos per user is 160, while 65% users have less than 160 clicked micro-videos, in the training set. The distribution of category-wise micro-video counts shows that almost every category has more or less micro-videos, which in part reveals the rationality of defining categories.

4.2 Compared Methods

We compare our proposed THACIL network against several baselines including the state-of-the-art methods.

Table 1: Statistics of the MicroVideo-1.7M dataset

	# Interactions	# Micro-videos	# Users	Sparsity
Training set	8,970,310	984,983	10,986	99.92%
Test set	3,767,309	719,897	10,986	99.97%

- BPR [23]: Bayesian personalized ranking is a pairwise ranking framework, which is to learn the relative ranking of two items for the same user. We implement BPR with the following settings: each user is embedded into a 128-dim vector, each micro-video is also embedded into a 128-dim vector (i.e. $d = 128$, with $d_c = d_f = 64$). All the embedding parameters are trained.
- LSTM [33]: Long- and short-term memory (LSTM) is an improved version of RNN with gated controls. We implement an LSTM network to model user's behaviors as sequence, whose output is combined with the item features to predict the click-through probability, just like the feedforward

network shown in Fig. 1 without attention. The number of parameters is configured to be identical (as possible) to our THACIL network.

- LSTM+Attention: On top of the implemented LSTM network, a vanilla attention mechanism is augmented, like the feedforward network in Fig. 1.
- NCF [13]: Neural collaborative filtering is a state-of-the-art method for deep network-based recommendation. It learns both user embedding and item embedding with a shallow network (element-wise product of user and item) and a deep network (concatenation of user and item followed by several feedforward layers). In our implementation, each user/item is embedded into a 128-dim vector, similar to the case of BPR.
- ATRank [35]: A very recent state-of-the-art method on modeling user behaviors as sequence, which features a specially designed attention mechanism. Our implementation of ATRank also employs the same feedfoward network in Fig. 1.

4.3 Implementation

All the methods including ours and the compared baselines are implemented with Tensorflow[6] and running on an NVIDIA GTX 1080Ti graphical processing unit. The following hyper-parameters are used for all the methods, if not specially noted.

- User and item embedding. For BPR and NCF, user is embedded into 128-dim vector and the user embedding is trained. For all the methods using item embedding, each item (micro-video) is embedded into a 128-dim vector, with 64-dim of category embedding and 64-dim of visual information embedding.
- Network structure. All the hidden layers have 128-dim output. For our THACIL, the number of micro-videos per user (n in Fig. 1) is set to 160, according to the statistics shown in Fig. 3a. The temporal block size (k in Fig. 1) is set to 20. For users having more items than 160, we just preserve as much as 160 items. For users having less items, we pad all-zero vectors to augment.
- Training. The batch size is by default 128. Adam optimizer [18] is adopted for training. Learning rate is 0.001. Weight decay with L_2 is set to 5×10^{-5}.

4.4 Evaluation Metrics

To evaluate the overall performance of micro-video click-through prediction by different methods, we adopt the widely used Area Under Curve (AUC) as the primary metric, which is defined as:

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{I}_u^+||\mathcal{I}_u^-|} \sum_{i \in \mathcal{I}_u^+} \sum_{j \in \mathcal{I}_u^-} \delta(\hat{s}_{u,i} > \hat{s}_{u,j}) \tag{11}$$

where $\hat{s}_{u,i}$ is the predicted score that a user $u \in \mathcal{U}$ may click a micro-video i in the test set, $\mathcal{U}$ is the set of all users, $\mathcal{I}_u^+$ and $\mathcal{I}_u^-$ consist of the micro-videos that the user u actually clicked and actually not clicked, respectively, $\delta(\cdot)$ is the indicator function.

In addition, the click-through prediction results can be used for generating recommendations. So we further evaluate the precision, recall, F-score, and Normalized Discounted Cumulative Gain (NDCG) of the top-K items of each user. Here, precision indicates the percentage of actually clicked items in the top-K list, recall is the percentage of retrieved clicked items, F-score is the harmonic average of precision and recall, and NDCG is calculated by dividing DCG by IDCG. The average results at different K values are reported.

Table 2: AUC results of different methods on the MicroVideo-1.7M test set

	BPR	LSTM	LSTM+Attention	ATRank	NCF	THACIL
AUC	0.583	0.641	0.654	0.660	0.672	**0.684**

4.5 Results

Table 2 presents the AUC results for measuring performance of different methods. According to the AUC results, we can observe that our method achieves significant improvement than all the other methods, which demonstrates the effectiveness of our proposed THACIL network for the task of micro-video click-through prediction. The most competitive baseline is NCF, so we perform a paired Student's t-test to compare our method and NCF, which shows the improvement in AUC is statistically significant ($p = 8.75 \times 10^{-7}$). For different views of the capabilities of different methods, we also report precision, recall, F-score, and NDCG results, as shown in Fig. 4. The relative ordering of the compared methods shows consistency between Table 2 and Fig. 4, indicating reliability of the results.

Since the focus of our THACIL network is to model user's historical behaviors as sequence, we can compare it with sequential modeling methods including LSTM, LSTM+Attention, and ATRank. It can be observed that LSTM+Attention performs better than LSTM, showing the benefit of the attention mechanism. In addition, ATRank performs only slightly better than LSTM+Attention, but it is worth noting that ATRank has higher computational efficiency than LSTM kind of methods. Our THACIL network outperforms all of them, owing to the advantage of capturing both short-term and long-term correlation within user behaviors. Our THACIL network is also more computationally efficient than LSTM and LSTM+Attention.

It is worth noting that NCF, not modeling user's behaviors as sequence, also achieves very good results in our experiments. An in-depth analysis of the results seems to indicate that NCF benefits a lot from the trained user embedding, which was not used in the other methods (except for BPR). Since the dataset has only 10,986 users, and the users in the training and test sets are exactly the same, trained user embedding shows advantage. However, the cost of user embedding is to store much more parameters (128-dim vector per user). Our THACIL network without trained user embedding achieves better performance than NCF, which further demonstrates the efficiency of our method. Also, note that BPR performs the worst in the compared methods, which partially shows the difficulty of the task.

[6] https://www.tensorflow.org

5 CONCLUSION

We have presented a recommendation algorithm for micro-videos, where we achieved the click-through prediction for micro-videos by modeling user's historical behaviors with the proposed THACIL network. Our network characterizes both short-term (i.e. within temporal block) and long-term (i.e. across temporal blocks) correlation within user behaviors, and profiles user interests at both coarse (category-level) and fine (item-level) granularities. We performed experiments on MicroVideo-1.7M, a new dataset coming from real data, and experimental results demonstrated the effectiveness of the THACIL network.

Personalized recommendation of micro-videos remains largely unexplored at many aspects. From the user side, in addition to user behavior modeling, demographic profiling and social cues could also be helpful in making recommendations. From the micro-video side, we may utilize much richer information including audio, color, and motion. We plan to investigate the integration of different modalities into a complete recommendation system.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under Grants 61772483, 61331017, 61390512, 61622211, 61472392 and 61620106009, and by the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*.

[2] 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 *WWW*. 895–904.

[3] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. 2012. Personalized video recommendation through tripartite graph propagation. In *MM*. 1133–1136.

[4] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *MM*. 898–907.

[5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*. 335–344.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *RecSys*. 191–198.

[7] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *MM*. 597–606.

[8] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension. In *COLING*. 1777–1786.

[9] 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 *RecSys*. 293–296.

[10] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5, 2 (2016), 99–113.

[11] Andrea Ferracani, Daniele Pezzatini, Marco Bertini, and Alberto Del Bimbo. 2016. Item-Based Video Recommendation: An Hybrid Approach considering Human Factors. In *ICMR*. 351–354.

[12] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A Unified Personalized Video Recommendation via Dynamic Recurrent Neural Networks. In *MM*. 127–135.

[13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. *ICLR* (2016).

[15] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *RecSys*. 241–248.

[16] Lei Huang and Bin Luo. 2017. Personalized Micro-Video Recommendation via Hierarchical User Interest Modeling. In *PCM*. Springer, 564–574.

[17] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. 2016. Real-time video recommendation exploration. In *SIGMOD*. 35–46.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*, 1,15.

[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR,abs:1411.2539* (2014).

[20] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual video recommendation by multimodal relevance and user feedback. *TOIS* 29, 2 (2011), 10.

[21] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 5528–5531.

[22] Jonghun Park, Sang-Jin Lee, Sung-Jun Lee, Kwanho Kim, Beom-Suk Chung, and Yong-Ki Lee. 2010. An online video recommendation framework using view based tag cloud aggregation. *IEEE Multimedia* 99, 1 (2010).

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*. 452–461.

[24] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *SIGIR*. 811–820.

[25] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. *AAAI* (2018).

[26] Yang Song, Ali Mamdouh Elkahky, and Xiaodong He. 2016. Multi-rate deep learning for temporal recommendation. In *SIGIR*. 909–912.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 6000–6010.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. *CVPR* (2018).

[29] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *WSDM*. 495–503.

[30] 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 *ICLR*. 2048–2057.

[31] Ming Yan, Jitao Sang, and Changsheng Xu. 2015. Unified youtube video recommendation via cross-network collaboration. In *ICMR*. 19–26.

[32] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. A dynamic recurrent model for next basket recommendation. In *SIGIR*. 729–732.

[33] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential Click Prediction for Sponsored Search with Recurrent Neural Networks.. In *AAAI*, Vol. 14. 1369–1375.

[34] Xiaojian Zhao, Guangda Li, Meng Wang, Jin Yuan, Zheng-Jun Zha, Zhoujun Li, and Tat-Seng Chua. 2011. Integrating rich information for video recommendation with multi-task rank aggregation. In *MM*. 1521–1524.

[35] Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiusi Chen, and Jun Gao. 2018. ATRank: An Attention-Based User Behavior Modeling Framework for Recommendation. *AAAI* (2018).

[36] Xiangmin Zhou, Lei Chen, Yanchun Zhang, Longbing Cao, Guangyan Huang, and Chen Wang. 2015. Online video recommendation in sharing community. In *SIGMOD*. 1645–1656.

[37] Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. 2013. Videotopic: Content-based video recommendation using a topic model. In *Multimedia (ISM) IEEE International Symposium on*. IEEE, 219–222.