Noname manuscript No.
(will be inserted by the editor)

Aspect-level Sentiment Capsule Network for Micro-Video Click-Through Rate Prediction

Yuqiang Han · Pan Gu · Wei Gao · Guandong Xu · Jian Wu

Received: date / Accepted: date

Abstract Micro-videos, a new form of videos that are constrained in duration, gain significant popularity in recent years. The volume and rate of online micro-videos urgently calls for effective recommendation algorithms to help users find their interested ones. Although some previous works have investigated how to model users' historical behaviors to predict the click-through rate of micro-videos, they are generally based on positive feedback only but overlook the negative which can help understand user preference at a finer granularity. The positive and negative feedback jointly imply the user's different sentiments on different aspects, where each aspect is one component of a micro-video such as *video_scene* and *video_subject*. To this end, we propose an **a**spect-level **s**entiment **cap**sule network(**ASCap**) for micro-video click-through rate prediction by aggregating both positive and negative feedback, with an attempt to make the prediction more explainable. More specifically, an aspect-specific gating mechanism is firstly utilized to extract the aspect-level features from the target micro-video and the user's positive and negative feedback. Then, in the following sentiment capsule network, the aspect-level features of the target micro-video are paired with those of positive and negative feedback respectively to identify their sentiments and form the sentiment capsules. Finally, the prediction layer is employed to calculate the overall click probability based on the sentiment capsules. Experimental results on two real-world micro-video datasets demonstrate that the proposed method significantly outperforms the state-of-the-art methods.

Yuqiang Han and Pan Gu contribute equally.

Jian Wu is the corresponding author.

Yuqiang Han, Wei Gao, Jian Wu
College of Computer Science and Technology, Zhejiang University, Hangzhou, China
E-mail: hyq2015@zju.edu.cn, E-mail: gw@zju.edu.cn, E-mail: wujian2000@zju.edu.cn

Pan Gu
College of Modern Science and Technology, China Jiliang University, Hangzhou, China
E-mail: L191800020@cjlu.edu.cn

Guandong Xu
Advanced Analytics Institute, University of Technology Sydney, Sydney, Australia
E-mail: Guandong.Xu@uts.edu.au

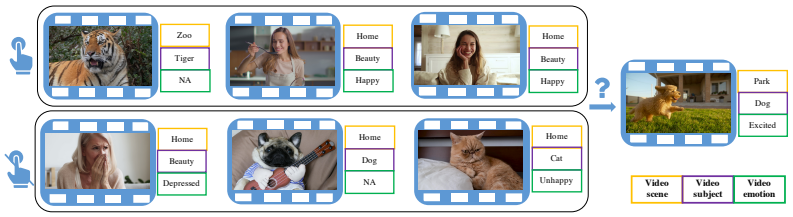


Fig. 1 An example of a user expressing different sentiments on different aspects of a micro-video.

Keywords Aspect-level Sentiment · Capsule Network · Micro-Video · Click-Through Rate Prediction

1 Introduction

Micro-videos, a new form of user-generated time-constrained(usually tens of seconds) videos, have become increasingly popular recently. Such bite-sized videos can be conveniently shot and shared by smartphones without the need for professional devices and skills, making the number of them growing exponentially in the online sharing platforms, such as Kuaishou[1], TikTok[2], and Vine[3]. Taking Kuaishou in China as an example, as of January 2020, there are more than 10 million daily uploaded new micro-videos and 300 million daily active users. Micro-videos also have many commercial potentials, such as brand promotion and online marketing. Hence, the micro-video sharing platforms must urgently build effective recommender systems to enhance the user experience, engagement, and retention.

In the last few years, personalized video recommendation has been well studied. The proposed algorithms can be categorized into three groups: collaborative filtering[3, 14], content-based filtering[7, 24, 42], and hybrid methods[4, 8, 35, 39]. However, micro-videos have a shorter length and low-quality descriptive text, and a user can interact with many relevant micro-videos in a short period, making the interaction sequence much longer. Thus, the micro-video recommendation is a more challenging task and has received increasing attention from the academic field in recent years. For example, Wei et al. [31] fused multi-modal features of micro-videos by graph convolution networks to better capture user preferences for personalized recommendation. Chen et al. [6] adopted hierarchical attention at an item- and category-level to model user behaviors for micro-video click-through rate prediction.

Although achieving promising results, the aforementioned methods are generally based on positive feedback only(*i.e.,* "click" behaviors) but overlook the negative, such as the "unclick" behaviors. Here, the "unclick" behavior in micro-video sharing platforms means that a user previews the thumbnail, yet no "click" behavior occurs, which may indicate what a user dislikes to a certain extent[25]. Some previous works have investigated to aggregate various types of feedback together to better understand user preference for recommendation[26, 40, 25, 19]. For example, in the most recent work, Li et al. [19] developed a unified framework to characterize

[1] `https://www.kuaishou.com`
[2] `https://www.tiktok.com`
[3] `https://www.vine.co.`

users' interested and uninterested micro-videos from "click" and "unclick" behavior sequences respectively, where each sequence is modeled by a temporal graph-based LSTM network. However, they modeled the user preference on the micro-videos at the item level and aggregated the interested and uninterested cues via a primitive weighted summation operation, which cannot well distinguish their impacts on user preference. Actually, a micro-video consists of several components, and the positive and negative feedback jointly imply the user's different sentiments on different components. As illustrated in Figure 1, each micro-video consists of three components {*video_scene, video_subject, video_emotion*}. From the "click" and "unclick" behavior sequences, we can see that the user expresses more positive attitude to beauty-related *video_subject* and positive *video_emotion*, more negative attitude to animals-related *video_subject* and negative *video_emotion*, and neutral attitude to *video_scene* which is present in both sequences. Whether the user will click the target micro-video depends on the mixture of all aspect sentiments. Hence, it is essential to model user preference at a finer level of granularity by taking both positive and negative feedback into consideration, which can make predictions with explanations.

To this end, in this paper, we propose an **a**spect-level **s**entiment **cap**sule network(ASCap) for micro-video click-through rate prediction by aggregating both positive and negative feedback, where each aspect is one component of a micro-video. Since there is no explicit labeled data indicating which aspects a user likes or dislikes, we exploit the implicit sentiment information provided by the "click" and "unclick" behaviors to guarantee the proper sentiment modeling. Intuitively, a user expresses positive sentiments to most aspects of a clicked micro-video, and negative attitudes to most aspects of an unclicked micro-video. Based on this, we delicately extract the aspect features with positive sentiment and those with negative sentiments from "click" and "unclick" behaviors, respectively. In addition, some aspect features which are frequently present in both "click" and "unclick" behaviors bear neutral sentiment. We argue that neutral sentiment also plays some role in understanding user preference, especially for some users, the positive and negative sentiment towards the micro-video is not obvious. To be more specific, ASCap firstly utilizes an aspect-specific gating mechanism to extract the aspect-level features from the target micro-video and the user's positive and negative behaviors. In the following sentiment capsule network, aspect-level features of the target micro-video pair with those of positive and negative feedback respectively to identify their sentiments and form the sentiment capsules. Finally, the prediction layer calculates the overall click probability based on all sentiment capsules. Experimental results on two real-world micro-video datasets demonstrate that ASCap significantly outperforms the state-of-the-art methods.

The main contributions of this work are summarized as follows:

- We propose an aspect-level sentiment capsule network based on gating mechanisms and capsule networks for micro-video click-through rate prediction at a finer level of granularity, with an attempt to make prediction more explainable.
- We introduce a capsule network-based architecture to derive and identify users' sentiment towards each aspect according to positive and negative feedback and further make improvements to the dynamic routing algorithm to make sure it can better identify the aspect sentiments.

– We perform extensive experiments on two public micro-video datasets to demonstrate that our proposed model significantly outperforms the state-of-the-art micro-video click-through rate prediction methods.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Section 3 presents the details of the proposed model. Section 4 reports the experimental settings and gives the analysis of results. Section 5 concludes the work and plans the future work.

2 Related Work

In this section, we briefly review the studies related to our research work: micro-video understanding and capsule networks.

2.1 Micro-video Understanding

In recent years, as a new form of user-generated content, micro-videos have gained increasing popularity in public. Meanwhile, micro-video content analysis has attracted extensive research efforts from the academic field, such as popularity prediction [5], venue categorization [36], hashtag recommendation [30], and micro-video recommendation [13, 6, 19, 23, 21, 20, 22]. For example, Chen et al. [5] proposed a transductive model to find the optimal latent common space, unifying and preserving information from different modalities, for predicting micro-video popularity. Zhang et al. [36] built a tree-guided multi-task multi-modal learning model to jointly learn a common space from multi-modalities and leverage the predefined Foursquare hierarchical structure to regularize the relatedness among venue categories. Wei et al. [30] leveraged GCNs to model the complicated interactions among <users, hashtags, micro-videos> and learn their representations to recommend hashtags for micro-videos.

Specific to the micro-video recommendation, the previous works mainly focus on multi-modal features and sequential user behaviors. For instance, Huang and Luo [13] proposed a personalized micro-video recommendation method using hierarchical user interest modeling based on multi-modal features, including visual, acoustic, textual, emotional, and social features. Ma et al. [23] simultaneously incorporated multi-source content data of items and multi-networks of users to learn user and item representations for recommendation. Liu et al. [21] proposed a user-video co-attention network to learn multi-modal information from both user and micro-video sides using an attention mechanism. Liu and Chen [20] employed self-attention to capture multi-modal features of different importance and made use of multi-head attention to learn users' preference from historical records to perform the next micro-video recommendation. Chen et al. [6] proposed a Temporal Hierarchical Attention at Category- and Item-Level (THACIL) network for user behavior modeling to predict the user's click-through rate of micro-videos.

Though promising performance improvement is achieved by these efforts, they are based on positive feedback only but overlook the negative. Li et al. [19] is the most relevant work with ours which jointly characterized users' interested and uninterested history records by temporal graph-based LSTM networks. However, they made the item-level analysis of user preference on the micro-videos and aggregated the interested and uninterested cues via a primitive weighted summation operation,

which cannot well distinguish their impacts on user preference. Therefore, we propose an aspect-level sentiment capsule network for micro-video click-through rate prediction at a finer level of granularity in consideration of a user emphasizing different aspects with different sentiments of a micro-video.

2.2 Capsule Networks

The capsule network is proposed as a hierarchical architecture to model the complex relations among latent features, which helps improve the representational limitations of CNNs and RNNs [11]. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part [28]. Sabour et al. [28] firstly applied vector-output capsules with dynamic routing to recognize highly overlapping digits and it achieved impressive performance. The dynamic routing mechanism ensures that low-level features can be selectively aggregated to form high-level ones. Then a number of methods were proposed to improve the performance of capsule networks [2, 12, 18, 29], such as Hinton et al. [12] proposed a new iterative routing procedure based on the EM algorithm, which measures the compatibility between matrix capsules by clustering them through Gaussian distributions.

In addition, capsule networks have also been applied to some NLP tasks, including text classification [38], relation extraction [37], zero-shot user intent detection [33], and multi-task learning [34]. Recently, some researchers have applied it to recommendation tasks [16, 17]. For example, Li et al. [16] utilized capsule networks to model the multiple interests from users' historical behaviors, where each output capsule encodes an interest. Li et al. [17] proposed a sentiment capsule architecture with a novel routing by a bi-agreement mechanism to identify the informative logic unit and the sentiment-based representations in user-item level for rating prediction. Inspired by [17], we exploit capsule network architectures to identify the aspect-level sentiments for micro-video click-through rate prediction.

3 Our Proposed Model

We propose an aspect-level sentiment capsule network for the micro-video click-through rate prediction. The overall architecture is shown in Figure 2. It mainly consists of four components: sequence encoding layer, aspect extraction layer, sentiment capsule layer, and prediction layer. The sequence encoding layer encodes the long user behavior sequence into a short one with a temporal window. The aspect extraction layer extracts aspect-related features with an aspect-specific gating mechanism. The sentiment capsule layer derives and identifies the sentiment of each aspect and outputs the sentiment capsules for the prediction layer. The prediction layer calculates click probability for each sentiment capsule and aggregates them to produce the overall click probability.

3.1 Problem Formulation

The task of micro-video click-through rate prediction is to build a model to estimate the probability of a user clicking on a specific micro-video. A user's historical records

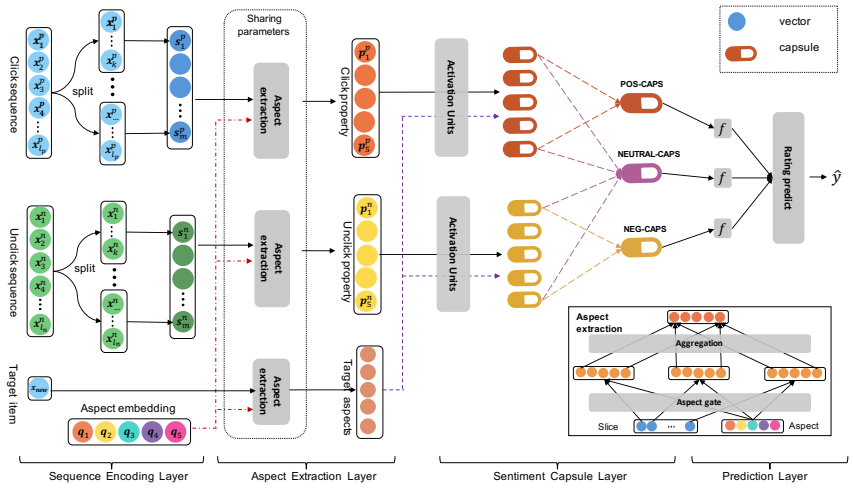


Fig. 2 The overall architecture of ASCap.

can be presented as an ordered sequence of micro-videos $\mathcal{U} = \{(u, x_j^p)\}_{j=1}^{l}$, where $p \in \{+, -\}$ respectively represents "click" and "unclick" behaviours, x_j is the identity of j-th micro-video, and l is the length of the sequence. The whole sequence can be segmented into two sub-sequences, namely "click" sequence $\mathcal{U}^+ = \{(u, x_j^+)\}_{j=1}^{l_+}$, and "unclick" sequence $\mathcal{U}^- = \{(u, x_j^-)\}_{j=1}^{l_-}$. As such, the micro-video click-through rate prediction problem can be formally defined as:

- **Input:** The user's "click" and "unclick" historical behavior sequences $\mathcal{U}^+$, $\mathcal{U}^-$, and the new micro-video x_{new}.
- **Output:** The click probability of the user u on the new micro-video x_{new}.

Considering that micro-videos are short of auxiliary information (tags, descriptions, *etc.*) in general [6], we use only visual information from the thumbnail to represent a micro-video in our work. This is reasonable because the thumbnail is the most representative snapshot, capturing the essence of a video and providing the first impression to the viewers. Besides, it can also alleviate the severe cold-start problem in micro-video click-through rate prediction. The mathematical notations used in this paper are summarized in Table 1.

3.2 Sequence Encoding

For a user, the click interaction sequence $\mathcal{U}^+$ can be represented as $\mathbf{X}^+ = [\mathbf{x}_1^+, \cdots, \mathbf{x}_{l_+}^+]$, where $\mathbf{x}_j^+ \in \mathbb{R}^d$ is the dense visual feature vector extracted from its cover picture, and d is the embedding size. As the shorter length of micro-videos, a user can interact with many videos in a short time, making the interaction sequence much longer than that of traditional videos. So we split the long sequence $\mathbf{X}^+$ into m blocks using a temporal window of width w based on the consideration that a user presses stable interest patterns in a short time, which means that the videos in a block mostly have the similar aspect features (*i.e.,* videos with the animals-related subject).

Table 1 Notations

Notations	Definitions and Descriptions
l_+	The length of the click sequence
l_-	The length of the unclick sequence
M	The number of aspects
w	The temporal window size
m	The number of blocks
$\mathcal{U}$	User interaction micro-video id sequence
$\mathcal{U}^+$	User clicked micro-video id sequence
$\mathcal{U}^-$	User unclicked micro-video id sequence
x_j^+	The j-th micro-video id in click sequence, $1 \leq j \leq l_+$
x_j^-	The j-th micro-video id in unclick sequence, $1 \leq j \leq l_-$
x_{new}	The given new micro-video id
$\mathbf{x}_j^+$	The j-th micro-video embedding in click sequence, $1 \leq j \leq l_+$
$\mathbf{x}_j^-$	The j-th micro-video embedding in unclick sequence, $1 \leq j \leq l_-$
$\mathbf{x}_{new}$	The new micro-video embedding
$\mathbf{X}^+$	The visual embedding of click sequence, $1 \leq j \leq l_+$
$\mathbf{X}^-$	The visual embedding of unclick sequence, $1 \leq j \leq l_-$
$\mathbf{s}_k^+$	The k-th block embedding of click sequence, $1 \leq k \leq m$
$\mathbf{s}_k^-$	The k-th block embedding of unclick sequence, $1 \leq k \leq m$
$\mathbf{q}_i$	The embedding of i-th aspect, $1 \leq i \leq M$
$\mathbf{p}_i^+$	The i-th aspect feature of click sequence, $1 \leq i \leq M$
$\mathbf{p}_i^-$	The i-th aspect feature of unclick sequence, $1 \leq i \leq M$
$\mathbf{p}_i^{new}$	The i-th aspect feature of the new micro-video, $1 \leq i \leq M$
$\mathbf{u}_i^+$	the i-th activation unit from click sequence, $1 \leq i \leq M$
$\mathbf{u}_i^-$	the i-th activation unit from unclick sequence, $1 \leq i \leq M$
$\mathbf{v}_{pos}$	the positive sentiment capsule
$\mathbf{v}_{neg}$	the negative sentiment capsule
$\mathbf{v}_{neu}$	the neutral sentiment capsule

In each block, we apply *sum pooling* operation to aggregate the local information as follows:

$$\mathbf{s}_k^+ = \sum_{j=wk}^{w(k+1)} \mathbf{x}_j^+ \tag{1}$$

In this way, $\mathbf{s}_k^+$ can encode the short-term preference within k-th block. Note that the last block will be padded to the same length as others if it contains less than w micro-videos. With the same procedure, we can get the "unclick" block representation $\mathbf{s}_k^-$. Although aggregating the videos in the same block may lose some information, the essential part will be preserved, which will be used to extract the aspect-level features and help identify the aspect sentiments of the given micro-video in the sentiment capsules network.

The block size is a hyper-parameter, which can be decided according to the characteristics of the micro-video platforms. In experiments, we set $w = 30$, which can achieve the best performance. It may be because a user can interact with about 30 micro-videos in a short time and express stable interest patterns on the micro-

video sharing platforms. The 30 micro-videos in a block are similar in some way, and we can accurately aggregate the information in each block, which benefits the model performance. The method is simple and proved efficient in our experiments, and is also a trade-off between efficiency and effectiveness. Besides, we have tried LSTM and attention methods to model the historical interactions in a sequential manner in each block, with no improvement. A possible reason is that the sequence data is noisy due to the rapid interaction behavior and presents no strong sequential pattern. Special structures may be needed to model such data in a sequence way. We leave it for future research.

3.3 Aspect Extraction

As discussed in Section 3.1, we present a micro-video using its visual embedding extracted from the thumbnail, and the fine-grained aspects are components of the video, such as *video_scene*, *video_subject*, and *video_emotion*. We use a shared aspect-specific gating mechanism to extract the aspect-level features relevant to the i-th aspect from k-th block as follows:

$$\mathbf{p}_{i,k}^{+} = \mathbf{s}_{k}^{+} \odot \sigma(\mathbf{W}_{i,1}\mathbf{s}_{k}^{+} + \mathbf{W}_{i,2}\mathbf{q}_{i} + \mathbf{b}_{i}) \tag{2}$$

where $\mathbf{W}_{i,1}$, $\mathbf{W}_{i,2} \in \mathbb{R}^{d\times d}$ are transform matrices relevant to i-th aspect, $\mathbf{b}_i \in \mathbb{R}^d$ is bias vector for the i-th aspect, σ is the *sigmoid* activation function and $\odot$ is the element-wise product operation. $\mathbf{s}_k^+$ is the k-th block embedding in "click" sequence. And $\mathbf{q}_i$ is the embedding of i-th aspect shared for all users, which is trained out by model optimization. The number of aspects M is a hyper-parameter and it can be determined by cross-validation. And then *average pooling* is applied to aggregate features with regard to the i-th aspect as follows:

$$\mathbf{p}_{i}^{+} = \frac{1}{m}\sum\nolimits_{k=1}^{m} \mathbf{p}_{i,k}^{+} \tag{3}$$

where m is the number of blocks. Finally, we can get M feature vectors $\mathbf{p}_i^+$ from click and $\mathbf{p}_i^-$ from unclick sequence, respectively. For the target micro-video $\mathbf{x}_{new}$, we can also get M aspect feature vectors $\mathbf{p}_i^{new}$, where $i \in \{1, \cdots, M\}$.

The aspect gate is shared by target micro-videos, "click" and "unclick" behaviors, which is to extract the features relevant to the specific aspect. For the severe cold-start problem, in some cases, the target micro-video may have new aspect features (*e.g.,* a new *video_style*) that are present in neither positive nor negative feedback, this aspect is dismissed and will not contribute to the final prediction. In our experiments, we set the number of aspects M as 5. More analysis can be found in Section 4.6.1. Note that, although only visual information is used in this work, multi-modal features can be easily employed if given, for example, each modality can be regarded as an aspect.

3.4 Sentiment Capsule

For a given micro-video, we predict a user's click probability by identifying his sentiments towards all aspects. To this end, we propose a capsule network with three

output capsules(*i.e.,* positive, negative, and neutral) to fulfill this task. That is to say, the aspects of a new micro-video will be "clustered" into three sentiment groups. Since there is no explicit labeled data to indicate which aspects a user likes or dislikes, we exploit the implicit sentiment information provided by the "click" and "unclick" behaviors to guarantee the proper sentiment modeling. Intuitively, a user expresses positive sentiments to most aspects of a clicked micro-video, and negative attitudes to most aspects of an unclicked micro-video. Based on this, we delicately extract the aspect features with positive sentiment and those with negative sentiments from "click" and "unclick" behaviors, respectively. In addition, some aspect features which are frequently present in both "click" and "unclick" behaviors bear neutral sentiment. We argue that neutral sentiment also plays some role in understanding user preference, especially for some users, the positive and negative sentiment towards the micro-video is not obvious. The feature extraction process is guided by the operations in sentiment transformation in Equation. 6 and the margin loss in Equation. 15.

Firstly, we pair aspect features of the given micro-video with those from "click" and "unclick" sequence respectively to form activation units. The activation units of "click" behaviors can be obtained as follows:

$$\mathbf{u}_i^+ = g(\mathbf{p}_i^{new} \odot \mathbf{p}_i^+) \tag{4}$$

where $\mathbf{p}_i^{new}$ is the i-th aspect feature vector of the new micro-video, $\mathbf{p}_i^+$ is the i-th aspect feature vector of the "click" sequence, $\odot$ is the element-wise product operation, and g is the non-linear squash function through the entire vector used in [28], which can be shown as:

$$\mathbf{u}_i = \frac{||\mathbf{h}_i||^2}{||\mathbf{h}_i||^2 + 1} \frac{\mathbf{h}_i}{||\mathbf{h}_i||} \tag{5}$$

where $||\cdot||$ denotes the length of a vector and $\mathbf{h}_i$ is the input capsule vector. It encodes the relevance between the aspects of new micro-video and corresponding ones from "click" sequence. In a similar way, we can get the activation units formed with aspects of new micro-video and corresponding ones from "unclick" sequence.

Then, the sentiment features are derived from activation units of "click" sequence as follows:

$$\hat{\mathbf{u}}_{s|i}^+ = \mathbf{H}_{i,s}^+ \mathbf{u}_i^+ \tag{6}$$

where $s \in \{pos, neu\}$, $\mathbf{H}_{i,s}^+ \in \mathbb{R}^{d \times d}$ is the sentiment transform matrix between i-th positive activation unit to s output capsule, which can derive the positive and neutral sentiment features from "click" sequence. The positive output capsule $\mathbf{v}_{pos}$ is a weighted sum over the corresponding sentiment features $\hat{\mathbf{u}}_{pos|i}^+$ as follows:

$$\mathbf{v}_{pos} = g(\sum_i^M c_{pos|i}^+ \hat{\mathbf{u}}_{pos|i}^+)) \tag{7}$$

where g is the non-linear squash function and $c_{pos|i}^+ \in [0, 1]$ is coupling coefficients as in Equation. 10 that are determined by the iterative dynamic routing process. Similarly, we can obtain the negative output capsule $\mathbf{v}_{neg}$ as follows:

$$\mathbf{v}_{neg} = g(\sum_i^M c_{neg|i}^- \hat{\mathbf{u}}_{neg|i}^-)) \tag{8}$$

The neutral capsule $\mathbf{v}_{neu}$ is a weighted sum over features derived from both "click" and "unclick" behaviors (neutral sentiments are present in both of them) as follows:

$$\mathbf{v}_{neu} = g(\frac{1}{2}\sum_i c^{+}_{neu|i}\hat{\mathbf{u}}^{+}_{neu|i} + \frac{1}{2}\sum_i c^{-}_{neu|i}\hat{\mathbf{u}}^{-}_{neu|i}) \tag{9}$$

To make the sentiment capsules more distinguishing, we make two improvements to guarantee that the dynamic routing method can better cluster the aspect sentiments. The improved routing-by-agreement algorithm is summarized in Algorithm 1.

- **Softmax with temperature.** We explore the softmax with temperature [10] in the place of standard softmax while updating connection strength between capsules in two layers:

$$\hat{c}^{p}_{s|i} = \frac{exp(b^{p}_{s|i}/\tau)}{\sum_i exp(b^{p}_{s|i}/\tau)}) \tag{10}$$

  where $p \in \{+,-\}$, $s \in \{pos, neg, neu\}$, $b^{p}_{s|i}$ is the logits of coupling coefficients that coupling input capsule i to output capsule s and generally initialized to 0, and τ is a temperature coefficient to tune. For a low temperature ($\tau \to 0^{+}$), an activation unit tends to concentrate on a single sentiment capsule, while for a high temperature ($\tau \to \infty$), it tends to concentrate on all sentiment capsules with nearly the same probability. In experiments, we set $\tau = 0.8$ to make each activation unit concentrate on a single sentiment, which can produce the best performance. It is because in this setting the model can make each activation unit concentrate on a specific sentiment capsule in a proper way, which contributes to the predictions.
- **Coefficients Amendment**. The length of activation unit vector indicates the probability of activation of the target aspect. So we attempt to employ it to iteratively amend the connection strength inspired by [38] as follows:

$$c^{p}_{s|i} = \hat{a}^{p}_{i} \cdot \hat{c}^{p}_{s|i} \tag{11}$$

  where $\hat{a}^{t}_{i} = ||\mathbf{u}^{t}_{i}||$ is the length of activation unit vector, $p \in \{+,-\}$, and $s \in \{pos, neg, neu\}$.

3.5 Prediction Layer

Given the sentiment capsule $\mathbf{v}_s$, we calculate the probability the user would click the given micro-video x_{new} as follows:

$$y_s = \mathbf{W}^{T}_{s}(tanh(\mathbf{H}_s\mathbf{v}_s + \mathbf{b}_{s,1})) + b_{s,2} \tag{12}$$

where $s \in \{pos, neg, neu\}$, $\mathbf{H}_s \in \mathbb{R}^{d\times d}$ and $\mathbf{W}^{T}_{s} \in \mathbb{R}^{d\times 1}$ are sentiment transform matrices, $\mathbf{b}_{s,1} \in \mathbb{R}^{d}$ is the bias vector, and $b_{s,2}$ is the scalar bias. Then the overall click probability can be calculated based on three probabilities as follows:

$$\hat{y} = \sigma(y_{pos} \cdot ||\mathbf{v}_{pos}|| + y_{neu} \cdot ||\mathbf{v}_{neu}|| + y_{neg} \cdot ||\mathbf{v}_{neg}|| + b_u) \tag{13}$$

where σ is the *sigmoid* function, b_u is the user bias, and $||\mathbf{v}||$ is the length of sentiment capsule vector, which indicates the confidence of prediction.

Algorithm 1: Improved Routing-by-Agreement Algorithm

Input: $\hat{\mathbf{u}}^p_{s|i}$ $\hat{a}^p_i$, r, l, where $p \in \{+,-\}$, $s \in \{pos, neg, neu\}$

Output: $\mathbf{v}_s$

Initialize the logits of coupling coefficients $b^p_{s|i} = 0$

for *r iteration* **do**

  for all capsule i in layer l and capsule s in layer $l+1$:

    $c^p_{s|i} = \hat{a}^p_i \cdot softmax(b^p_{s|i}/\tau)$

  `/* Update the sentiment capsule */`

  $\mathbf{v}_{pos} = g(\sum_i c^+_{pos|i} \hat{\mathbf{u}}^+_{pos|i})$

  $\mathbf{v}_{neg} = g(\sum_i c^-_{neg|i} \hat{\mathbf{u}}^-_{neg|i})$

  $\mathbf{v}_{neu} = g(\frac{1}{2}\sum_i c^+_{neu|i} \hat{\mathbf{u}}^+_{neu|i} + \frac{1}{2}\sum_i c^-_{neu|i} \hat{\mathbf{u}}^-_{neu|i})$

  for all capsule i in layer l and capsule s in layer $l+1$:

    $b^p_{s|i} = b^p_{s|i} + \hat{\mathbf{u}}^p_{s|i} \cdot \mathbf{v}_s$

end

return $\mathbf{v}_s$

3.6 Model Optimization

We use *sigmoid cross-entropy loss* to guide the parameter learning for model optimization:

$$L(y, \hat{y}) = -(y \log \sigma(\hat{y}) + (1-y) \log(1 - \sigma(\hat{y})) \tag{14}$$

where $y \in \{0, 1\}$ is the ground truth indicating whether the user clicks the target micro-video, and σ is the *sigmoid* function.

To insure the capsules can correctly reflect fine-grained sentiments, we use a separate *margin loss* [28] for sentiment capsule network as a regularization:

$$\begin{aligned} L_{stm} = \frac{1}{|\mathcal{O}|} \sum (&\max(0, \epsilon - ||\mathbf{v}_s||) \\ &+ \lambda \max(0, ||\mathbf{v}_{\neg s}|| - 1 + \epsilon)) \\ &+ \lambda \max(0, ||\mathbf{v}_{neu}|| - 1 + \epsilon)) \end{aligned} \tag{15}$$

where $\mathcal{O}$ denotes the set of observed user-item pairs, $\epsilon = 0.8$ and $\lambda = 0.5$, empirically. When the ground truth $y = 1$, $\mathbf{v}_s = \mathbf{v}_{pos}$; otherwise, $\mathbf{v}_s = \mathbf{v}_{neg}$. $\neg s$ denotes the opposite of sentiment s.

In addition, we introduce a *disagreement regularization* to explicitly encourage the diversity among multiple aspects as follows:

$$L_{asp} = -\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{\mathbf{q}_i}{||\mathbf{q}_i||} \frac{\mathbf{q}_j}{||\mathbf{q}_j||} \tag{16}$$

where M is the number of aspects.

So, the final loss can be represented as follows:

$$L = L(y, \hat{y}) + \lambda_s L_{stm} + \lambda_a L_{asp} \tag{17}$$

where the regularization weights $\lambda_s = 0.1$ and $\lambda_a = 0.1$ in our experiments. We use Adam[15] for parameters update in an end-to-end fashion.

4 Experiments

In this section, we conduct experiments on two real-world micro-video datasets to verify the effectiveness of our proposed method for micro-video click-through rate prediction.

4.1 Dataset

The experimental datasets are from two popular micro-video sharing platforms.

- **Kuaishou-3.2M**. The original version of the dataset is released by the Kuaishou Competition[4] in *ChinaMM* 2018 conference. In our experiments, we used the sampled dataset constructed by [19], which consists of randomly selected $10,000$ users and their $13,661,383$ interactions with $3,239,534$ micro-videos. The interaction types include "click", "unclick", "like", and "follow". Each micro-video's visual features are represented by a $2,048$-d embedding vector of the thumbnail. Each user's historical interactions are sorted in chronological order. We set the first 80% of a user's historical interactions as the training set and the rest 20% as the test set following the method in [19]. Furthermore, in the training set, we take the first 90% interactions of each sequence for training and the last 10% for validation.
- **MicroVideo-1.7M**. This dataset is constructed by [6], which consists of $10,986$ users and their $12,737,619$ interactions with $1,704,880$ micro-videos. The interaction types include "click" and "unclick". Each micro-video is represented by a 512-d visual embedding vector of its thumbnail. Each user's historical interactions are sorted in chronological order. We divide the micro-videos into two disjoint sets and divide the interactions according to micro-videos into two sets, one for training and the other for the test following the method in [6]. Moreover, for each sequence in the training set, the first 90% interactions are used for training and the last 10% for validation.

The statistics of two experimental datasets are shown in Table 2. **Particularly, in both datasets, "unclick" behavior means the user does not click the micro-video after previewing its thumbnail, which is recorded by the micro-video sharing platforms.** In addition, we also adopted the Principal Component Analysis (PCA)[32] to reduce the micro-video's visual embedding to 64 dimension following the method in [19].

4.2 Evaluation Metrics

To evaluate the effectiveness of different methods, we use Area Under Curve (AUC) as the primary metric, which is also widely used in other related works.

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{I}_u^+||\mathcal{I}_u^-|} \sum_{i \in \mathcal{I}_u^+} \sum_{j \in \mathcal{I}_u^-} \delta(\hat{y}_{u,v_i} - \hat{y}_{u,v_j}) \tag{18}$$

[4] https://www.kuaishou.com/activity/uimc

Table 2 Statistics of the two datasets.

Statistics	Kuaishou-3.2M	MicroVideo-1.7M
# Users	10,000	10,986
# Items	3,239,534	1,704,880
# Interactions	13,661,383	12,737,619
# Interaction types	4	2
# Interactions per user	1366.14	1159.44
# Interactions per item	4.28	7.47
# Clicked items per user	277	218
# Training - Interactions	9,833,373	8,508,406
# Validation - Interactions	1,097,719	900,917
# Test - Interactions	2,730,291	3,767,309

where $\hat{y}_{u,v_i}$ is the predicted probability that a user u may click a given micro-video i in the test set, $\mathcal{U}$ is the set of all users, $\mathcal{I}_u^+$ and $\mathcal{I}_u^-$ respectively represent the set of micro-videos that the user u actually clicked and unclicked, and $\delta(\cdot)$ is the indicator function.

In addition, we also employ Precision, Recall, and F-score to further evaluate the model performance from different angles. Given the top-K recommendation list computed based on the predicted click probability, P@K indicates the percentage of actually clicked micro-videos in the list, R@K means the percentage of retrieved clicked micro-videos, and F@K is the harmonic average of precision and recall. The statistical significance test is conducted by performing the paired *t-test*.

4.3 Comparison Methods

To demonstrate the effectiveness of our model for micro-video click-through rate prediction, we compare it to the following representative and state-of-the-art methods.

- **BPR**[27]: Bayesian personalized ranking is a popular pairwise ranking framework, which is to model the relative preferences of users by a pairwise loss function.
- **CNN-R**: This method utilizes the 1-D CNN structure to model user behavior sequences. Explicitly, the kernel size varies from 1 to 10, and the number of filters under each kernel size is 32.
- **LSTM-R**: This model utilizes the LSTM network to extract the sequential pattern from the user's behavior sequence. It generates the preference representation by feeding the hidden states in each step into a fully connected layer, followed by an MLP layer to predict the click probability.
- **ATRank**[41]: This is an attention-based user behavior modeling framework, which projects all types of behaviors into multiple latent semantic spaces and makes the influence among different behaviors via self-attention.
- **NCF**[9]: This is a collaborative filtering based recommendation framework, which learns latent features of users and items with a shallow network, and leverages a multi-layer perceptron to learn the user-item interaction function.
- **THACIL**[6]: This is a personalized micro-video recommendation method by modeling user's historical behaviors, which leverages category- and item-level

attention mechanisms to model the diverse and fine-grained interests respectively, and adopts forward multi-head self-attention to capture the long-term correlation within user behaviors.
- **ALPINE**[19]: This is a personalized micro-video recommendation method considering the diverse and dynamic interest, multi-level interest, and true negative samples. It utilizes a temporal graph-based LSTM network to model users' dynamic and diverse interests from click sequence, and capture uninterested information from the true negative sample. Beyond that, it introduces a user matrix to enhance user interest modeling by incorporating multiple types of interactions.

Note that, for CNN-R, LSTM-R, BPR, and NCF, we fed the user representation and the target micro-video embedding into an MLP layer to predict the final click probability.

4.4 Parameters Setup

For all compared baseline methods, we set the parameters as used in [19], which are selected by applying grid search based on the setting strategies reported in their papers. For Kuaishou-3.2M, we use the 64-d visual embedding to represent the micro-video. For MicroVideo-1.7M, we take the concatenation of the 64-d visual embedding and the 64-d category embedding as micro-video embedding for baselines, where the category embedding is trained, but only 64-d visual embedding is used in our method. For BPR and NCF, user embedding is initialized as a 128-d vector and is learned out. The maximum length of users' historical sequence is set to 300. If it has more items than 300, we truncate it to 300; otherwise, we pad all-zero vectors that are masked in the network to augment. In addition, for our proposed model, the temporal window size is set to 30, iteration of dynamic routing r is set to 2, the temperature τ in softmax is 0.8, the number of aspects M is 5, and the aspect embedding size and sentiment capsule vector size are both 64. The parameters are optimized using Adam with an initial learning rate 0.001 and mini-batch size 512. All weight matrices are initialized by sampling from the normal distribution $N(0, 0.1^2)$, and all biases are set to zeros. The final performances of all methods are reported over five runs with the same hyper-parameters to exclude the impact of random parameter initialization. The model is defined and trained in TensorFlow[1] on a GeForce GTX 1080 Ti GPU. The code will be released soon.

4.5 Performance Evaluation

A summary of the results of all methods under the metric of AUC, P@50, R@50, and F@50 over the two datasets are reported in Table 3. Several observations can be made from the results:

(1) The sequential methods LSTM-R and CNN-R surpass the non-sequential method BPR, demonstrating that the sequential information plays a role in capturing users' interests. Moreover, the self-attention based models, i.e., ATRank and THACIL, outperform CNN-R and LSTM-R, revealing that more delicate structures are needed to accurately capture the user preference from a long user behavior sequence, so as to focus on the key interest information. ALPINE further

Table 3 Performance comparison over two datasets. The best and the second best results are highlighted in boldface and underline respectively. † indicates the ASCap significantly outperforms the best baseline over AUC at $p = 0.05$ level by t-test.

Methods	Kuaishou-3.2M				MicroVideo-1.7M			
	AUC	P@50	R@50	F@50	AUC	P@50	R@50	F@50
BPR	0.595	0.290	0.387	0.331	0.583	0.241	0.181	0.206
LSTM-R	0.713	0.316	0.420	0.360	0.641	0.277	0.205	0.236
CNN-R	0.719	0.312	0.413	0.356	0.650	0.287	0.214	0.245
ATRank	0.722	0.322	0.426	0.367	0.660	0.297	0.221	0.253
NCF	0.724	0.320	0.420	0.364	0.672	0.316	0.225	0.262
THACIL	0.727	0.325	0.429	0.369	0.684	**0.324**	0.234	0.269
ALPINE	<u>0.739</u>	<u>0.331</u>	<u>0.436</u>	<u>0.376</u>	<u>0.713</u>	0.300	<u>0.460</u>	<u>0.362</u>
ASCap	**0.742**†	**0.338**	**0.443**	**0.383**	**0.725**†	0.314	**0.473**	**0.377**
Improv.	0.41%	2.11%	1.61%	1.86%	1.68%	-3.09%	2.83%	4.14%
ASCap-RA	0.740	0.336	0.441	0.381	0.724	0.313	0.472	0.376

improves the recommendation performance by characterizing the user's uninterested cues in addition to the interested and multi-level interest information, suggesting that different types of feedback can help better understand user preference.

(2) Although not modeling the sequential information from the user's behavior sequence, NCF achieves better performance than BPR and sequential models. It mainly contributes to its high expressiveness by fusing the linear MF and nonlinear MLP models. Therefore NCF can better model the relationship between users and items and produce better user embeddings.

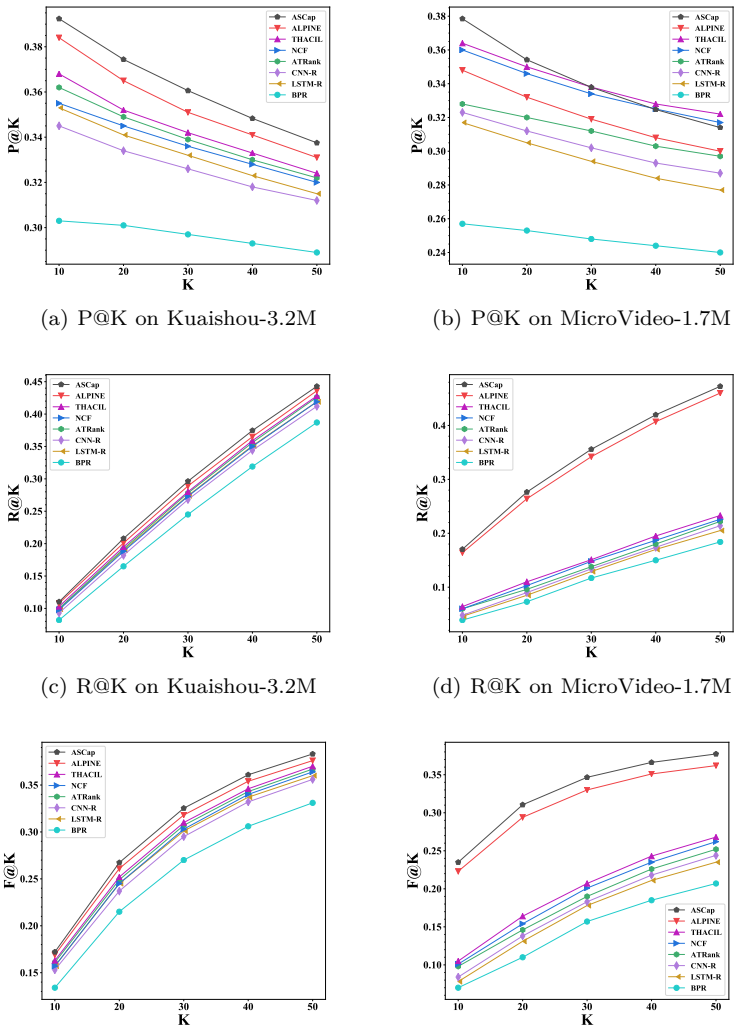
(3) Our proposed model ASCap achieves the best performance generally, which suggests its efficacy for the task of micro-video click-through rate prediction. Particularly, ASCap presents improvements over ALPINE, which captures the interested and uninterested information from "click" and "unclick" historical interactions at item-level, verifying that the fine-grained user preference model can bring more performance improvements. Note that ASCap performs a little worse than the best baseline at P@50 over MicroVideo-1.7M, which may be because of only visual information used in our method.

To justify the robustness of our proposed model from different views, we also report the performances of all methods by varying the number of returned micro-videos K from 10 to 50 in Figure 3. From the results we can see that:

(1) The performance of all methods over Recall and F value upgrades with the number of returned items K increasing. At the same time, our proposed model ASCap consistently performs better than all baselines.

(2) Besides, the performance of all methods over Precision value degrades as increasing the number of returned items K. Meanwhile, our model can consistently outperform others on Kuaishou-3.2M and achieve competitive results on MicroVideo-1.7M.

All observations verify the robustness and capabilities of ASCap for the task of micro-video click-through prediction.



(a) P@K on Kuaishou-3.2M (b) P@K on MicroVideo-1.7M

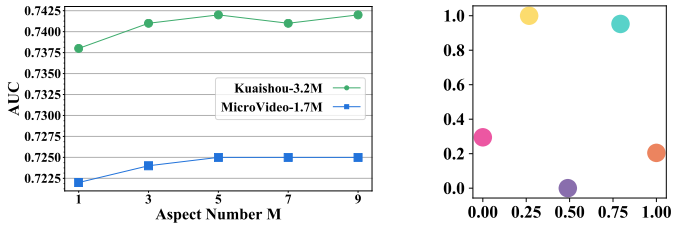
(c) R@K on Kuaishou-3.2M (d) R@K on MicroVideo-1.7M

(e) F@K on Kuaishou-3.2M (f) F@K on MicroVideo-1.7M

Fig. 3 Performance versus the number of returned micro-videos K.

4.6 Analysis of ASCap

We further make a detailed analysis of two primary components of ASCap (i.e., dynamic routing and aspect extraction) to give an in-depth understanding of the working process over on MicroVideo-1.7M (results of Kuaishou-3.2M showing similar patterns are omitted). We fix the other parameters to the values described in Section 4.4.



(a) Influence of Aspect Number M. (b) Aspect Embedding.

Fig. 4 Analysis results of aspects.

4.6.1 Influence of Aspect Number

The M value specifies the number of aspects extracted from each micro-video, which can be determined by cross-validation. To investigate the influence of M, we report the performance patterns of ASCap by tuning M amongst {1, 3, 5, 7, 9} in Figure 4(a). We can see that ASCap with multiple aspects achieves better performance than a single one, suggesting it is beneficial to model user preference in a fine-grained manner. Given the performance variation is small when $M >= 5$, we choose to use $M = 5$ in our experiments on two datasets.

To have an in-depth understanding of the learned aspects, we visualize the embedding of 5 aspects trained in our experiment in Figure 4(b), where each colored point represents one aspect. We can clearly see that all aspects separate from each other, ensuring that they can be used to extract different features from a micro-video, such as the *video_scene*, *video_subject*, and *video_emotion* illustrated in Figure 1.

4.6.2 Influence of Dynamic Routing

We report the results of model with standard route-by-agreement algorithm as used in [28] in the last row of Table 3. Although ASCap-RA achieves competitive performance, ASCap algorithm performs better, demonstrating the efficacy of *softmax* with temperature and coefficients amendment in the improved routing-by-agreement algorithm.

Furthermore, we randomly select four examples to explore the details of sentiment capsules. And we also choose one aspect from each case to show how it is routed to the corresponding capsule. From the results shown in Table 4, we can see that:

(1) For the first example, the length of positive sentiment capsule is larger than negative, i.e., $||\mathbf{v}_{pos}|| > ||\mathbf{v}_{neg}||$, so the positive sentiment dominants the negative, and the predicted click probability $\hat{y} = 0.687$ is consistent with ground truth $y = 1$. As to the chosen aspect, the coefficients of the positive activation unit $c^{+}_{pos|1} = 0.435$ is much larger than $c^{+}_{neu|1} = 0.095$, and those of the negative activation units $c^{-}_{neg|1} = 0.053$ and $c^{-}_{neu|1} = 0.045$ are very small, meaning that the user express positive sentiment towards this aspect of the target micro-video and this aspect is routed to the positive capsule.

(2) For the second example, the length of negative sentiment capsule is larger than positive, i.e., $||\mathbf{v}_{neg}|| > ||\mathbf{v}_{pos}||$, so the negative sentiment dominants, and the

Table 4 Example study of two users from MicroVideo-1.7M.

y	$\hat{y}$	$\|\|\mathbf{v}_{pos}\|\|$	$\|\|\mathbf{v}_{neg}\|\|$	$\|\|\mathbf{v}_{neu}\|\|$	$c^{+}_{pos\|1}$	$c^{+}_{neu\|1}$	$c^{-}_{neg\|1}$	$c^{-}_{neu\|1}$
1	0.687	0.729	0.591	0.056	0.435	0.095	0.053	0.045
0	0.098	0.382	0.829	0.230	0.093	0.103	0.502	0.110
0	0.056	0.468	0.755	0.572	0.188	0.483	0.087	0.133
0	0.142	0.322	0.735	0.258	0.057	0.054	0.050	0.061

predicted click probability $\hat{y} = 0.098$ is consistent with ground truth $y = 0$. As to the chosen aspect, the coefficients of the negative activation unit $c^{-}_{neg|1} = 0.502$ is much larger than $c^{-}_{neu|1} = 0.110$ and the coefficients of the positive activation units $c^{+}_{pos|1} = 0.093$ and $c^{+}_{neu|1} = 0.103$ are small, expressing that the user express negative sentiment towards this aspect of the target micro-video and this aspect is routed to the negative capsule.

(3) In the third example, although the overall negative sentiments dominate the positive, specific to the chosen aspect, we can see that $c^{+}_{neu|1} > c^{+}_{pos|1}$ and $c^{-}_{neu|1} > c^{-}_{neg|1}$, indicating that the features of the aspect are present in both positive and negative feedback. So it indicates that the user expresses neutral sentiment towards this aspect and it will be routed to the neutral capsule.

(4) In the fourth case where the overall negative sentiments dominate the positive, the 4 coefficients of the chosen aspect are all very small, meaning that this aspect contains new features such as new *video_style* that is present in neither "click" nor "unclick" behaviors. Therefore, the model can extract few features from it and it will be dismissed when making predictions.

The above analysis results suggest that the sentiment capsule network can effectively derive aspect features with sentiments, cluster the aspects of a given micro-video into three sentiment groups, and make accurate click-through rate prediction.

5 Conclusion and future work

In this paper, we proposed an aspect-level sentiment capsule network based on gating mechanisms and capsule networks for micro-video click-through rate prediction at a finer level of granularity by aggregating positive and negative feedback. We introduced a capsule network-based architecture to identify a user' sentiment towards each aspect and further made improvements to the dynamic routing algorithm to better identify the aspect sentiments. We performed extensive experiments on two public micro-video datasets to demonstrate that our proposed model ASCap significantly outperforms the state-of-the-art micro-video click-through rate prediction methods.

In future work, we plan to extend the model with multi-modal features, such as acoustic and textual modality, to better model user preference. Also, we plan to investigate more user-video interaction types, e.g., "like" and "follow", to better model users' sentiment towards micro-videos.

Acknowledgements This work was partially supported by the Zhejiang University Education Foundation under grants No. K18-511120-004, No. K17-511120-017, and No. K17-518051-021, the National Natural Science Foundation of China under grant No. 61672453, the National key R&D program sub project "large scale cross-modality medical knowledge management" under grant No. 2018AAA0102100.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. (2016) Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp 265–283
2. Bahadori MT (2018) Spectral capsule networks
3. Baluja S, Seth R, Sivakumar D, Jing Y, Yagnik J, Kumar S, Ravichandran D, Aly M (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, pp 895–904
4. Chen B, Wang J, Huang Q, Mei T (2012) Personalized video recommendation through tripartite graph propagation. In: Proceedings of the 20th ACM international conference on Multimedia, pp 1133–1136
5. Chen J, Song X, Nie L, Wang X, Zhang H, Chua TS (2016) Micro tells macro: predicting the popularity of micro-videos via a transductive model. In: Proceedings of the 24th ACM international conference on Multimedia, pp 898–907
6. Chen X, Liu D, Zha ZJ, Zhou W, Xiong Z, Li Y (2018) Temporal hierarchical attention at category-and item-level for micro-video click-through prediction. In: Proceedings of the 26th ACM international conference on Multimedia, pp 1146–1153
7. Cui P, Wang Z, Su Z (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, pp 597–606
8. Ferracani A, Pezzatini D, Bertini M, Del Bimbo A (2016) Item-based video recommendation: An hybrid approach considering human factors. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp 351–354
9. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS (2017) Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web, pp 173–182
10. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:150302531
11. Hinton GE, Krizhevsky A, Wang SD (2011) Transforming auto-encoders. In: International conference on artificial neural networks, Springer, pp 44–51
12. Hinton GE, Sabour S, Frosst N (2018) Matrix capsules with em routing

13. Huang L, Luo B (2017) Personalized micro-video recommendation via hierarchical user interest modeling. In: Pacific Rim Conference on Multimedia, Springer, pp 564–574
14. Huang Y, Cui B, Jiang J, Hong K, Zhang W, Xie Y (2016) Real-time video recommendation exploration. In: Proceedings of the 2016 International Conference on Management of Data, pp 35–46
15. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
16. Li C, Liu Z, Wu M, Xu Y, Zhao H, Huang P, Kang G, Chen Q, Li W, Lee DL (2019) Multi-interest network with dynamic routing for recommendation at tmall. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp 2615–2623
17. Li C, Quan C, Peng L, Qi Y, Deng Y, Wu L (2019) A capsule network for recommendation and explaining what you like and dislike. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 275–284
18. Li H, Guo X, DaiWanli Ouyang B, Wang X (2018) Neural network encapsulation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 252–267
19. Li Y, Liu M, Yin J, Cui C, Xu XS, Nie L (2019) Routing micro-videos via a temporal graph-guided recommendation system. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1464–1472
20. Liu S, Chen Z (2019) Sequential behavior modeling for next micro-video recommendation with collaborative transformer. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 460–465
21. Liu S, Chen Z, Liu H, Hu X (2019) User-video co-attention network for personalized micro-video recommendation. In: The World Wide Web Conference, pp 3020–3026
22. Ma J, Li G, Zhong M, Zhao X, Zhu L, Li X (2018) Lga: latent genre aware micro-video recommendation on social media. Multimedia Tools and Applications 77(3):2991–3008
23. Ma J, Wen J, Zhong M, Chen W, Zhou X, Indulska J (2019) Multi-source multi-net micro-video recommendation with hidden item category discovery. In: International Conference on Database Systems for Advanced Applications, Springer, pp 384–400
24. Mei T, Yang B, Hua XS, Li S (2011) Contextual video recommendation by multimodal relevance and user feedback. ACM Transactions on Information Systems (TOIS) 29(2):1–24
25. Ouyang W, Zhang X, Li L, Zou H, Xing X, Liu Z, Du Y (2019) Deep spatio-temporal neural networks for click-through rate prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2078–2086
26. Peska L, Vojtas P (2013) Negative implicit feedback in e-commerce recommender systems. In: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, pp 1–4
27. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2012) Bpr: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:12052618
28. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866

29. Wang D, Liu Q (2018) An optimization view on dynamic routing between capsules
30. Wei Y, Cheng Z, Yu X, Zhao Z, Zhu L, Nie L (2019) Personalized hashtag recommendation for micro-videos. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1446–1454
31. Wei Y, Wang X, Nie L, He X, Hong R, Chua TS (2019) Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1437–1445
32. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics and intelligent laboratory systems 2(1-3):37–52
33. Xia C, Zhang C, Yan X, Chang Y, Yu PS (2018) Zero-shot user intent detection via capsule neural networks. arXiv preprint arXiv:180900385
34. Xiao L, Zhang H, Chen W, Wang Y, Jin Y (2018) Mcapsnet: Capsule network for text with multi-task learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 4565–4574
35. Yan M, Sang J, Xu C (2015) Unified youtube video recommendation via cross-network collaboration. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp 19–26
36. Zhang J, Nie L, Wang X, He X, Huang X, Chua TS (2016) Shorter-is-better: Venue category estimation from micro-video. In: Proceedings of the 24th ACM international conference on Multimedia, pp 1415–1424
37. Zhang X, Li P, Jia W, Zhao H (2019) Multi-labeled relation extraction with attentive capsule network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 7484–7491
38. Zhao W, Ye J, Yang M, Lei Z, Zhang S, Zhao Z (2018) Investigating capsule networks with dynamic routing for text classification. arXiv preprint arXiv:180400538
39. Zhao X, Li G, Wang M, Yuan J, Zha ZJ, Li Z, Chua TS (2011) Integrating rich information for video recommendation with multi-task rank aggregation. In: Proceedings of the 19th ACM international conference on Multimedia, pp 1521–1524
40. Zhao X, Zhang L, Ding Z, Xia L, Tang J, Yin D (2018) Recommendations with negative feedback via pairwise deep reinforcement learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1040–1048
41. Zhou C, Bai J, Song J, Liu X, Zhao Z, Chen X, Gao J (2018) Atrank: An attention-based user behavior modeling framework for recommendation. In: Thirty-Second AAAI Conference on Artificial Intelligence
42. Zhou X, Chen L, Zhang Y, Cao L, Huang G, Wang C (2015) Online video recommendation in sharing community. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp 1645–1656