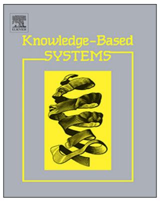
Knowledge-Based Systems 203 (2020) 106114

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities

Da Cao [a], Lianhai Miao [a], Huigui Rong [a,*], Zheng Qin [a], Liqiang Nie [b]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, PR China
[b] School of Computer Science and Technology, Shandong University, Qingdao, Shandong, 266000, PR China

ARTICLE INFO

Article history:
Received 1 July 2019
Received in revised form 28 May 2020
Accepted 4 June 2020
Available online 8 June 2020

Keywords:
Micro-videos
Hashtag recommendation
Multiple modalities
Deep neural network

ABSTRACT

Due to the short attention span and instant gratification phenomenon, micro-videos are growing exponentially while gaining more and more concerns. Yet the sheer number of micro-videos leads to severe information overload issues, making it difficult for users to identify their desired micro-videos. The hashtag, mainly utilized in the domain of the microblog or the image, is the indicator or the core idea of the target content and can be applied to various information retrieval scenarios (e.g., search, browse, and categorization). So far, however, little attention has been paid to perform the hashtag recommendation for micro-videos via harnessing multiple modalities.

In this article, we devise a neural network-based solution, LOGO (short for "mu**L**ti-m**O**dal-based hashta**G** rec**O**mmendation"), to recommend hashtags for micro-videos by utilizing multiple modalities. The proposed LOGO approach first represents each modality as the combination of sequential units in it, weighted by the attention mechanism. In this way, the sequential and attentive features are captured simultaneously. After that, the LOGO integrates the representations of all modalities via a multi-view representation learning framework, which projects the representations into a common space under the restriction of the modality similarity. Ultimately, the LOGO feed the projections of three modalities in the common space and the embeddings of hashtags into a customized neural collaborative filtering framework to perform the hashtag recommendation. Extensive experiments on the scope of both overall performance comparison and micro-level analyses have well-justified the effectiveness and rationality of our proposed approach.

1. Introduction

Boosted by the proliferation of the advanced multimedia technologies and portable devices, micro-video platforms, such as Vine[1] and Snapchat,[2] have emerged rapidly in recent years. The micro-video, as an extension of traditional textual and graphical media, has gradually become a new medium for the public to access information and perform social behaviors. Different from traditional long videos, micro-videos are generally recorded and shared within 6 to 15 seconds (The maximum length of the micro-video on Vine is restrict to 6 seconds, while that of Snapchat is 15 s.). Due to the limited time span, micro-videos are conveniently shot and instantly shared, leading to its widespread and huge volumes. Therefore, it is highly desired to devise some sophisticated methods for us to locate our truly desired micro-videos.

Hashtags, widely applied in the domain of microblogs [1], images [2], and news [3], are a concise representation of the concrete material and the key source for search engines to target the desired content. Considering the popular microblog platform Twitter[3] as an example, the hashtag is prefixed with the symbol of # and is usually utilized to mark keywords or key topics within a microblog. The tagging service can benefit users in searching and browsing their desired microblogs. Therefore, it is straightforward to come up with an idea of applying tagging service to the domain of micro-video for the information retrieval application. The micro-video tagging service is beneficial to the stakeholders of micro-video ecosystems. For users, the hashtags facilitate them to search and locate their desired micro-videos. For providers, concise and concrete hashtags can improve the probability of their micro-videos being discovered. For platforms, the hashtags are helpful for the fine-grained categorization. However, most users do not have the habit of tagging their micro-videos.

* Corresponding author.
E-mail addresses: caoda0721@gmail.com (D. Cao), lianhaimiao@gmail.com (L. Miao), ronghg@hnu.edu.cn (H. Rong), zqin@hnu.edu.cn (Z. Qin), nieliqiang@gmail.com (L. Nie).

[1] https://vine.co.

[2] https://www.snapchat.com.

[3] https://twitter.com.

https://doi.org/10.1016/j.knosys.2020.106114
0950-7051/© 2020 Published by Elsevier B.V.

According to the statistics on our collected micro-videos, as revealed in Fig. 1, up to 65% of the 584,876 micro-videos have no hashtags and the majority of hashtags occur less than 5 times. Thereby, how to devise a framework to recommend hashtags for micro-videos is a valuable and urgent research issue.

Despite its value and urgency, recommending hashtags for micro-videos remains an unaddress research issue. Specifically, micro-videos are encoded with the sequential structure within visual, acoustic, textual modalities (i.e., a set of ordered image frames, a sequential audio clips with successive amplitude of wave, and a series of semantically and syntactically correlated words). Meanwhile, the importance of different units (i.e., image frames, audio clips, and words) in each modality varies a lot. Therefore, how to capture the sequential structure and endow the units with separate weights in each modality is important to enhance the representation for micro-videos. In addition, micro-videos, exhibiting triple-heterogeneities, consist of visual, acoustic, and textual modalities. Different modalities depict the intrinsic content of micro-videos from different angles. How to effectively leverage the valuable information in the multiple modalities and seamlessly sew them up is a highly challenging problem we face.

To address the aforementioned challenges, we present an end-to-end solution, LOGO (short for "mu**L**ti-m**O**dal-based hashta**G** rec**O**mmendation"), comprising of three consecutive components. To be more specific, we first utilize three parallel Long Short-Term Memory Networks (LSTMs) [4,5] to model the sequential structure for units in each modality. The attention mechanism [6] is adopted to apply unit-aware weights to sequential units in each modality. In the second stage, to handle the triple-heterogeneous data structure, the outputs of the three parallel LSTMs are projected into a common space via three mapping functions. The projections are then regularized via the restriction of the modality similarity within each micro-video and each hashtag. Lastly, the projections of three modalities in the common space and the embeddings of hashtags are cast into a customized neural collaborative filtering (NCF) framework, regarding the hashtag recommendation task as a binary classification problem. By conducting experiments on our constructed real-world dataset, our proposed approach is demonstrated to yield significant gains as compared with other state-of-the-art competitors.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first work that attempts to recommend hashtags for micro-videos with a multi-modal learning framework by jointly utilizing visual, acoustic, and textual modalities.
- We develop a novel solution, LOGO, to improve the hashtag recommendation for micro-videos by jointly considering the sequential structure and multi-modal fusion. This work promotes the study of both sequential structure learning and heterogeneous data fusion.
- We have released our self-constructed dataset and our implementation to facilitate the research community for further exploration.[4]

2. Related work

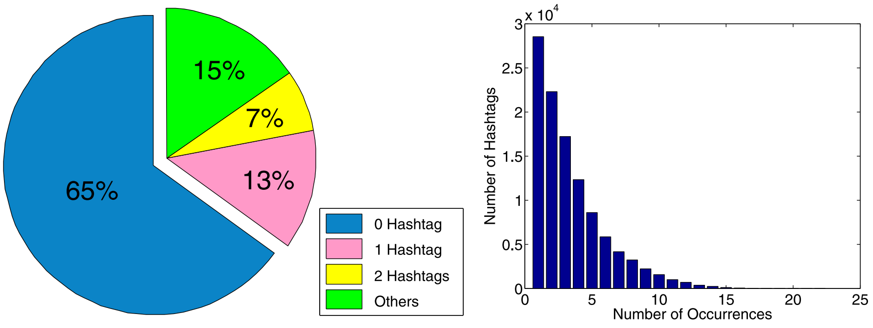
2.1. Multi-modal fusion

A wide range of approaches have been proposed for fusing image and text, and have been applied to various scenarios, such as sentiment analysis [7,8], cross-modal retrieval [9,10], and jointly modeling content and link [11,12]. In addition to image and text, audio is also an important ingredient for us to understand the content of micro-videos [13,14]. Motivated by the multi-view learning, a common space from multiple modalities (i.e., visual, acoustic, and textual modalities) is jointly learnt in the work of [15] and [16], and is further applied to the venue category estimation and popularity prediction [17] for micro-videos. For better micro-video understanding, the work of [18] characterizes and jointly models the sparseness and multiple sequential structures among multiple modalities. A deep transfer model is presented in [19] to transfer the external sound knowledge to strengthen the low-quality acoustic modality in micro-videos. In this paper, we focused on recommending hashtags by jointly considering the visual, acoustic, and textual information, which is a novel subject of multi-modal fusion.

2.2. Hashtag recommendation

Hashtags have received great attention in recent years and have been widely applied in various scenarios, such as popularity prediction [20], immersive search [21], and event detection [22]. When it comes to the application of recommendation [23–26], prior efforts in hashtag recommendation can be divided into three categories — probabilistic graph-based, neural network-based, and the hybrid of the two. Probabilistic graph-based approaches exploit the hashtags by modeling the hashtag generating process via the probability theory. In the work of [27], the hashtag recommendation task is modeled as a translation process from the content to hashtags. The social influence is presented in [28] to enhance the hashtag recommendation performance. A generative method is presented in [29] to incorporate the textual and visual information simultaneously. In [30], different types of hashtags are regarded with different distributions and then incorporated into an topical translation model. Different from probabilistic graph-based approaches, neural network-based methods explore the hashtag recommendation task by utilizing the specialities of deep neural networks, such as the attention mechanism and the sequential learning. To incorporate the trigger words, an attention-based CNN is investigated in [31] to perform the hashtag recommendation task. A co-attention network is proposed in [32] to recommend hashtags for multi-modal tweets by incorporating both textual and visual information. The work of [33] regards the hashtag recommendation as a classification task and proposes a novel recurrent neural network model to perform the hashtag recommendation for tweets. In order to take advantage of both the probability theory and neural network, some hybrid algorithms are developed. To learn deep item representations for the tag recommendation, the work of [34] jointly performs the deep representation learning and relational learning in a principled way under a probabilistic framework. Observing that hashtags indicate the primary topics of microblog posts, an attention-based LSTM model in [35] incorporates the topic modeling into the LSTM architecture through an attention mechanism. Besides, in the domain of micro-videos, a novel tag refinement approach is proposed in [36], which learns from multiple public data sources with manually labeled tags. Meanwhile, to provide a personalized hashtag recommendation for micro-videos, the work of [37] leverages recently advanced graph convolution network techniques to model the complicate interactions among <users, hashtags, micro-videos> and learn their representations. Although pioneer works have worked hard to perform the hashtag recommendation, none of them attempts to solve this issue by jointly utilizing visual, acoustic, and textual modalities.

[4] https://tagrec.wixsite.com/logo.



(a) Distribution of micro-videos (b) Distribution of hashtags

Fig. 1. Statistics on our collected micro-videos and their contained hashtags. Fig. 1(a) shows the distribution of micro-videos w.r.t. their contained hashtags and Fig. 1(b) reveals the distribution of hashtags w.r.t. their occurrences.

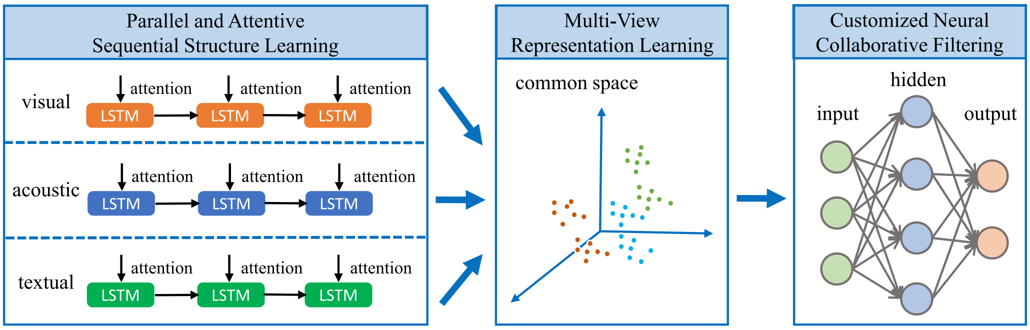


Fig. 2. Graphical representation of our proposed LOGO framework.

3. Our proposed framework

The framework of LOGO is illustrated in Fig. 2. Generally speaking, our proposed LOGO model is comprised of three components: (1) the parallel and attentive sequential structure learning component represents each modality of a micro-video as a fixed length of vector; (2) the multi-view representation learning component maps the vector representations of multiple modalities into a common space with the same length; and (3) the interaction learning with NCF casts the mappings of modalities in the common space and the embeddings of hashtags into a customized NCF framework. Specifically, we first present the notation and formulate the hashtag recommendation problem to be solved (Section 3.1). After that, we elaborate the three key ingredients of our proposed framework (Sections 3.2–3.4). We finally discuss the training process in detail (Section 3.5).

3.1. Notation and problem formulation

We use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to represent matrices and vectors, respectively. Meanwhile, we employ non-bold letters (e.g., x) to denote scalars, squiggle letters (e.g., $\mathcal{X}$) to sets, and Greek letters (e.g., λ) to parameters. If not clarified, all vectors are in column forms.

Suppose there are I micro-videos $\mathcal{X} = \{x_i\}_{i=1}^{I}$ and J hashtags $\mathcal{Y} = \{y_j\}_{j=1}^{J}$. For each micro-video $x \in \mathcal{X}$, we pre-segment it into three modalities $x = \{x^v, x^a, x^t\}$, whereinto the superscripts v, a, and t respectively represent the visual, acoustic, and textual modalities. Furthermore, we denote $x_i^m \in \mathbb{R}^{D_m}$ as a D_m-dimensional feature vector for the mth modality of the ith micro-video. Then, given a target micro-video x_i, our task is to recommend a list of hashtags to which the micro-video may be relevant.

3.2. Parallel and attentive sequential structure learning

The framework of attentive sequential structure learning is illustrated in Fig. 3, composed of the feature extraction from modalities, three parallel LSTMs, and an attention-based pooling. We first detail the sequential units partition and feature extraction from the visual, acoustic, and textual modalities. And then, we introduce the parallel LSTMs for sequential learning in multiple modalities. Lastly, we elaborate the attention mechanism applied in the output of parallel LSTMs.

3.2.1. Feature learning from modalities

In this part, we introduce the features that we extract from visual, acoustic, and textual modalities, respectively.

Visual Modality. Due to the short length of micro-videos, we are able to utilize a few key frames to represent the visual content of the whole micro-video. Specifically, we employ the tool of FFmpeg[5] to extract the key frames of micro-videos with the time gap of every 0.5 s. For each frame, the AlexNet model [38] is adopted to extract the visual features. To provide a robust initialization for recognizing semantics, the AlexNet model is pre-trained on a set of 1.2 million clean images from ILSVRC-2012.[6] Finally, we obtain 12 frames for each micro-video and a 4096 dimensional vector representation for each frame.

Acoustic Modality. As an important complement to the visual modality, the acoustic modality is especially useful in the cases where the visual content is too diverse or contain insufficient information. To extract the acoustic units, we partition each audio channel into 6 clips with the uniform length. Thereafter, we perform a spectrogram with a 46 ms window and 50% overlap

[5] https://www.ffmpeg.org/tool.

[6] http://www.image-net.org/challenges/LSVRC/2012.

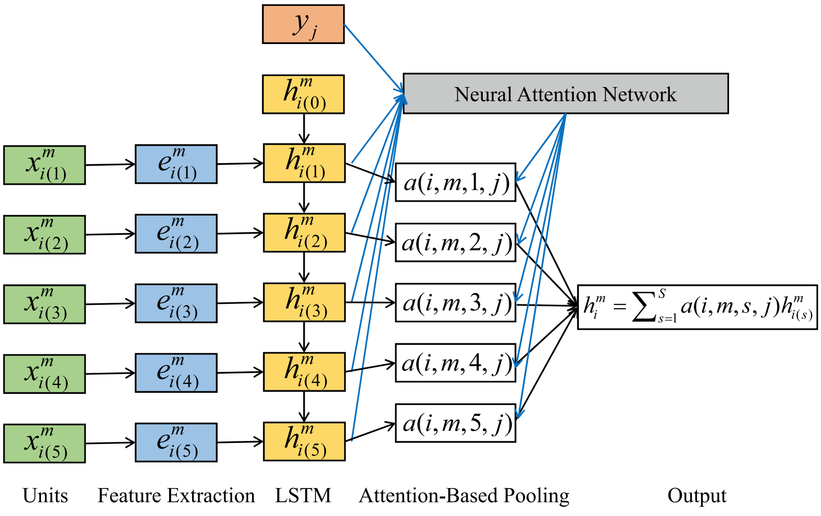


Fig. 3. The visualization of parallel and attentive sequential structure learning.

via Librosa,[7] generating 512 dimensional features for each audio clip.

Textual Modality. Textual descriptions are generally utilized as the key source for hashtag recommendation [31]. The textual modality is naturally divided into separated words. Word embedding (more accurately, word2vec [39]) is employed to generate the vector representations for words. The word2vec tool takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. To reduce the burden of model training, we directly obtain 300 dimensional features for each word via the prominent large-scale training results released by Google.[8]

3.2.2. *Parallel LSTMs*

The features extracted from sequential units in each modality are represented as $\{\mathbf{e}^m_{i(1)}, \ldots, \mathbf{e}^m_{i(S)}\}$, where $\mathbf{e}^m_{i(s)}$ denotes the vector representation for the sth unit in the mth modality of the ith micro-video. The features are then fed into parallel LSTMs to capture the sequential structure in visual, acoustic, and textual modalities, respectively. At the sth time step, LSTM takes the vector $\mathbf{e}^m_{i(s)}$, hidden state vector $\mathbf{h}^m_{i(s-1)}$, and memory cell vector $\mathbf{c}^m_{i(s-1)}$ as input, and updates $\mathbf{h}^m_{i(s)}$ and $\mathbf{c}^m_{i(s)}$ as follows:

$$\begin{cases} \mathbf{i}^m_{i(s)} = \sigma(\mathbf{W}^m_i \mathbf{e}^m_{i(s)} + \mathbf{U}^m_i \mathbf{h}^m_{i(s-1)} + \mathbf{b}^m_i) \\ \mathbf{f}^m_{i(s)} = \sigma(\mathbf{W}^m_f \mathbf{e}^m_{i(s)} + \mathbf{U}^m_f \mathbf{h}^m_{i(s-1)} + \mathbf{b}^m_f) \\ \mathbf{o}^m_{i(s)} = \sigma(\mathbf{W}^m_o \mathbf{e}^m_{i(s)} + \mathbf{U}^m_o \mathbf{h}^m_{i(s-1)} + \mathbf{b}^m_o) \\ \mathbf{g}^m_{i(s)} = \tanh(\mathbf{W}^m_g \mathbf{e}^m_{i(s)} + \mathbf{U}^m_g \mathbf{h}^m_{i(s-1)} + \mathbf{b}^m_g), \\ \mathbf{c}^m_{i(s)} = \mathbf{f}^m_{i(s)} \odot \mathbf{c}^m_{i(s-1)} + \mathbf{i}^m_{i(s)} \odot \mathbf{g}^m_{i(s)} \\ \mathbf{h}^m_{i(s)} = \mathbf{o}^m_{i(s)} \odot \tanh(\mathbf{c}^m_{i(s)}) \\ \quad m \in \{v, a, t\} \end{cases} \qquad (1)$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ are the element-wise sigmoid and hyperbolic tangent functions, respectively; $\odot$ is the element-wise multiplication operator; $\mathbf{i}^m_{i(s)}$, $\mathbf{f}^m_{i(s)}$, and $\mathbf{o}^m_{i(s)}$ are respectively treated as input, forget, and output gates for the mth modality. At $s = 1$, the initializations of $\mathbf{h}^m_{i(0)}$ and $\mathbf{c}^m_{i(0)}$ are set to zero vectors. In summary, parameters of the LSTMs are $\mathbf{W}^m_l$, $\mathbf{U}^m_l$, and $\mathbf{b}^m_l$ for $l \in \{i, f, o, g\}$.

3.2.3. *Attention-based pooling*

The output of a LSTM is a sequence of vectors $\{\mathbf{h}^m_{i(1)}, \ldots, \mathbf{h}^m_{i(S)}\}$. To facilitate the follow-up training process, it is essential to convert the set of variable-length vectors to a fix-length vector, termed as pooling.

In fact, there are some efforts made to design the pooling strategies. The most commonly used pooling operations in neural networks are average pooling and max pooling. Average pooling is a straightforward approach, taking the average hidden states from all time steps as the final representation for the sequence, namely, $\mathbf{h}^m_i = \frac{1}{S}\sum_{s=1}^{S} \mathbf{h}^m_{i(s)}$. Max pooling is similar to the average pooling, taking the element-wise max of all hidden states, namely, $\mathbf{h}^m_i = \max(\mathbf{h}^m_{i(1)}, \ldots, \mathbf{h}^m_{i(S)})$. However, we argue that such simple operations are unable to assign proper credits for individual units in a sequence. Another strategy is to extract the last hidden state $\mathbf{h}^m_{i(S)}$ as the representation for the whole sequence as implemented in [18], and we termed it as last pooling. However, the long term dependency problem occurs in this approach because the last hidden state is forced to remember inputs that are many steps away.

To overcome the weaknesses of the aforementioned pooling methods, we present a novel attention-based pooling, stemming from the neural machine translation [6], and extending to a general setting to adapt to our scenario. The attention-based pooling denotes a sequence by a weighted sum of the vector representations of all time steps. Particularly, the weights are relevant to the content of each time step, making it possible to dynamically adjust the weights of steps according to their importance. Mathematically, we parameterize $\alpha(i, m, s, j)$ as a neural attention network with the hidden state vector $\mathbf{h}^m_{i(s)}$ and the hashtag embedding $\mathbf{y}_j$ as the input:

$$\begin{cases} a(i, m, s, j) = (\mathbf{w}_m)^T \mathrm{ReLU}(\mathbf{W}^m_h \mathbf{h}^m_{i(s)} + \mathbf{W}^m_y \mathbf{y}_j + \mathbf{b}^m_a) \\ \alpha(i, m, s, j) = \mathrm{softmax}(a(i, m, s, j)) \\ \qquad = \dfrac{\exp a(i, m, s, j)}{\sum_{s=1}^{S} \exp a(i, m, s, j)} \end{cases}, \qquad (2)$$

where $\mathbf{W}^m_h$ and $\mathbf{W}^m_y$ are weight matrices of the attention network that convert the hidden state vector and the hashtag embedding to the hidden layer, and $\mathbf{b}^m_a$ is the bias vector of the hidden layer. We employ ReLU as the activation function for the hidden layer, and then project it to a score $a(i, m, s, j)$ with a weight vector $\mathbf{w}_m$. We normalize the scores with a softmax function, which is a common practice in previous efforts [6,40]. Lastly, the

[7] https://github.com/bmcfee/librosa.

[8] https://code.google.com/archive/p/word2vec.

representation of the sequence is an attention-based combination of all hidden states:

$$\mathbf{h}_i^m = \sum_{s=1}^{S} \alpha(i, m, s, j)\mathbf{h}_{i(s)}^m. \tag{3}$$

3.3. Multi-view representation learning

The outputs of parallel LSTMs are vector representations for visual, acoustic, and textual modalities with different lengths. How to seamlessly sew them up is the key to discover the intrinsic content of a micro-video. Traditional multi-modal fusion approaches generally concatenate the vector representations of all modalities into one vector, and then feed the concatenation result into a machine learning model. Another choice is to apply the machine learning method to all the modalities individually, and then integrate their results. However, we argue that these two approaches employ simple assembling strategies while ignore the crucial correlation among modalities. Therefore, we propose a multi-view representation learning method, treating each modality as an independent view and projecting the outputs of LSTMs into a common space. In the common space, the vector representations of visual, acoustic, and textual modalities are forced to be related under the constraint of the modality similarity. To map the varying lengths of vector representations of multiple modalities to the common space with the same length, we utilize three parallel mapping functions as follows:

$$\begin{cases} \widetilde{\mathbf{x}}_i^v = f_v(\mathbf{h}_i^v) \\ \widetilde{\mathbf{x}}_i^a = f_a(\mathbf{h}_i^a), \\ \widetilde{\mathbf{x}}_i^t = f_t(\mathbf{h}_i^t) \end{cases} \tag{4}$$

where $\widetilde{\mathbf{x}}_i^v$, $\widetilde{\mathbf{x}}_i^a$, and $\widetilde{\mathbf{x}}_i^t \in \mathbb{R}^c$ are the visual, acoustic, and textual embeddings in the common space, respectively; c is the dimension of the embeddings in the learned common space; $\mathbf{h}_i^v$, $\mathbf{h}_i^a$, and $\mathbf{h}_i^t$ are the vector representations derived from the attention-based parallel LSTMs; $f_v(\cdot)$, $f_a(\cdot)$, and $f_t(\cdot)$ are three mapping functions, which are approximated by the neural networks with the structure of multi-layer perceptron. To regularize the mapping functions, namely, the parameters in the neural networks, we propose a joint optimization framework to force (1) the modalities of each micro-video to be close; and (2) the modalities tagged with the same hashtag to be close. Based upon this, the objective function of the multi-view representation learning is stated as:

$$\begin{aligned} \mathcal{L}_{view} = & \sum_{i=1}^{I} \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}, m' \neq m} \|\widetilde{\mathbf{x}}_i^m - \widetilde{\mathbf{x}}_i^{m'}\|^2 \\ & + \lambda \sum_{j=1}^{J} \sum_{m \in \mathcal{M}} \sum_{r \in \mathcal{K}_j} \sum_{r' \in \mathcal{K}_j, r' \neq r} \|\widetilde{\mathbf{x}}_r^m - \widetilde{\mathbf{x}}_{r'}^m\|^2, \end{aligned} \tag{5}$$

where $\mathcal{M}$ denotes the set of modality indicators, namely, $\mathcal{M} = \{v, a, t\}$; $\mathcal{K}_j$ represents indexes of micro-videos tagged with the jth hashtag, namely, $\mathcal{K}_j = \{k_{j,1}, k_{j,2}, \ldots, k_{j,*}\}$ and $x_{k_{j,*}} \in \mathcal{X}$; λ balances the modality similarity restriction of the micro-video and hashtag. As such, the visual, acoustic, and textual embeddings can enhance each other for the better micro-video representation, which is crucial to the follow-up interaction learning process.

3.4. Customized neural collaborative filtering

NCF is neural network-based solution for item recommendation [41]. Its idea is to feed user embedding and item embedding into a dedicated multi-layer neural network (which needs to be customized) to learn the interaction function from data. As neural networks have a strong ability to fit the data, the NCF framework is more generalizable than the traditional matrix factorization (MF) model, which simply applies a data-independent inner product function as the interaction function. As such, we opt for the NCF framework to perform an end-to-end learning on both embeddings (that represent multiple modalities of micro-videos and hashtags) and interaction functions (that predict the interactions between micro-videos and hashtags).

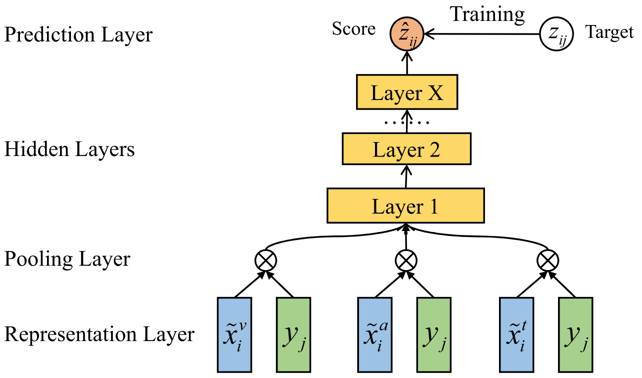


Fig. 4. Interaction learning based on neural networks.

Fig. 4 illustrates our interaction learning framework. Given an interaction between a micro-video and a hashtag (x_i, y_j), the representation layer first returns the embedding vectors for the micro-video and the hashtag, respectively. The former comes from the result of the multi-view representation learning (see details in Section 3.3), and the latter is generated from the neural network. Thereafter, the embeddings of both micro-videos and hashtags are cast into the pooling layer and hidden layers to achieve the prediction score.

Pooling Layer. Given an interaction between a micro-video and a hashtag (x_i, y_j), the pooling layer first performs element-wise products on the embeddings of each modality $\widetilde{\mathbf{x}}_i^m$ and the hashtag $\mathbf{y}_j$, and then concatenates their results:

$$\mathbf{p}_0 = \varphi_{pooling}(\widetilde{\mathbf{x}}_i^v, \widetilde{\mathbf{x}}_i^a, \widetilde{\mathbf{x}}_i^t, \mathbf{y}_j) = \begin{bmatrix} \widetilde{\mathbf{x}}_i^v \odot \mathbf{y}_j \\ \widetilde{\mathbf{x}}_i^a \odot \mathbf{y}_j \\ \widetilde{\mathbf{x}}_i^t \odot \mathbf{y}_j \end{bmatrix}, \tag{6}$$

where $\odot$ denotes the element-wise product of two vectors. The principles are two-fold. (1) The element-wise product inherits MF [42], which employs the multiplication on each embedding dimension to model the interaction between two embedding vectors. Element-wise product has been demonstrated to be highly effective in the feature interaction modeling with the low-level neural architecture [43]. And (2) the multiple modalities of micro-videos are of equal importance and have been generated as vector representations with the same length, the same as the element-wise products. To equally integrate the element-wise products, we utilize concatenation as the integration strategy.

Hidden Layers. Above the pooling layer is a stack of fully connected layers, which enable the model to capture the non-linear and high-order correlations between the micro-videos and hashtags. Formally, the hidden layers are defined as:

$$\begin{cases} \mathbf{p}_1 = \mathrm{ReLU}(\mathbf{W}_1\mathbf{p}_0 + \mathbf{b}_1) \\ \mathbf{p}_2 = \mathrm{ReLU}(\mathbf{W}_2\mathbf{p}_1 + \mathbf{b}_2) \\ \qquad \vdots \\ \mathbf{p}_l = \mathrm{ReLU}(\mathbf{W}_l\mathbf{p}_{l-1} + \mathbf{b}_l) \end{cases}, \tag{7}$$

where $\mathbf{W}_l$, $\mathbf{b}_l$, and $\mathbf{p}_l$ denote the weight matrix, bias vector and output neurons of the lth hidden layer, respectively. ReLU is used as the non-linear activation function, empirically verified to work effectively for tower-structure feed-forward neural networks [41, 44]. Also, the tower structure is utilized for hidden layers. Finally, the output of the last hidden layer $\mathbf{p}_l$ is transformed to a prediction score via the following formula:

$$\hat{z}_{ij} = \text{Sigmoid}({\mathbf{w}_p}^T \mathbf{p}_l), \tag{8}$$

where $\mathbf{w}_p$ denotes the weights of the prediction layer. Sigmoid is utilized to regularize the prediction score to the range of 0 to 1.

To learn the parameters of the neural network, we need to specify an objective function. We regard the hashtag recommendation task as a binary classification problem. In particular, an observed interaction between a micro-video and a hashtag is assigned to a target value of 1, or otherwise 0. We optimize the pointwise log loss, as implemented in [41], which forces the prediction score $\hat{z}_{ij}$ to close to the target z_{ij}:

$$\begin{aligned}\mathcal{L}_{network} &= -\sum_{(i,j)\in\mathcal{Q}} \log \hat{z}_{ij} - \sum_{(i,j)\in\mathcal{Q}^-} \log(1-\hat{z}_{ij}) \\ &= -\sum_{(i,j)\in\mathcal{Q}\cup\mathcal{Q}^-} z_{ij}\log \hat{z}_{ij} + (1-z_{ij})\log(1-\hat{z}_{ij}),\end{aligned} \tag{9}$$

where $\mathcal{Q}$ denotes the set of observed interactions, and $\mathcal{Q}^-$ denotes the set of negative instances.

To regularize the parameters in our proposed LOGO framework, we ultimately define our objective function by jointly regularizing the multi-view representation learning and the interaction learning with neural networks:

$$\mathcal{L} = \mathcal{L}_{view} + \mathcal{L}_{network}. \tag{10}$$

3.5. Optimization

3.5.1. Training scheme

For the training from scratch, we sample a fixed number of negative interactions (i.e., micro-video and hashtag pairs) to pair each observed interaction. Thereafter, we employ the stochastic gradient descent (SGD) to train our model in a mini-batch mode and update the corresponding model parameters by using the back propagation strategy. Specifically, we initially sample a batch of micro-videos and hashtags, and take a gradient step to optimize $\mathcal{L}_{view}$. And then we sample a batch of training interactions to optimize $\mathcal{L}_{network}$. To speed up the convergence rate of SGD, various modifications to the update rule have been explored, such as momentum, adagrad, and adadelta.

3.5.2. Dropout

Although neural network-based methods have the strong representation ability, they are easy to overfit the training data and cannot be effectively generalized to testing scenarios. To prevent deep neural networks from overfitting, we resort to an effective solution — dropout. The idea is to drop part of neurons during the training, and then only part of the model parameters will be updated. In our LOGO model, we propose to employ dropout on the pooling layer by randomly dropping ρ percent of the $\mathbf{p}_0$ vector. Moreover, we also apply dropout on the hidden layer of the neural attention network and the hidden layers of our interaction learning component. It is worth noting that dropout is only used during the training stage, and has to be disabled during the prediction phase.

4. Experiments

In this section, we conducted extensive experiments on our constructed dataset to answer the following four research questions:

- **RQ1** How does our proposed LOGO approach perform as compared with other state-of-the-art competitors?
- **RQ2** How do LOGO and other baselines behave for popular, relative popular, and unpopular hashtags?
- **RQ3** How do different predefined settings (e.g., the number of negative samples, the dropout ratio ρ, and the tradeoff parameter λ) affect LOGO?
- **RQ4** Are multiple modalities equally important? How does LOGO perform with different modality combinations?

4.1. Experimental settings

4.1.1. Micro-video dataset

To validate our work, we integrated two public micro-video datasets released by the previous studies [15,16]. After removing the duplicate micro-videos, 584,876 micro-videos are retained. Furthermore, we processed the dataset by retaining micro-videos with 3 modalities (i.e., the visual, acoustic, and textual modalities) and hashtag information, and hashtags occurred at least 5 times. Based upon these criteria, we ultimately obtained 108,661 micro-videos and 10,677 hashtags. On average, each micro-video has 2.65 hashtags, and each hashtag occurs 26.97 times.

The original dataset was further divided into three disjoint sets, with 80%, 10%, and 10% randomly selected micro-videos and their corresponding interactions for training, validation and testing, respectively. The validation set is leveraged to tune hyper-parameters and the final performance comparison is conducted on the testing set. Meanwhile, as reveal in Eq. (9), for each positive interaction, we paired it with a fixed number of negative interactions. Specifically, we randomly sampled 6 hashtags (details will be illustrated in Section 4.4) that the micro-video has never been marked to pair with each positive interaction. Meanwhile, each negative instance is assigned to the target value of 0.

4.1.2. Evaluation protocols

We followed the *leave-one-out* evaluation protocol, as implemented in [41,45,46], to evaluate the ranking performance. Since it is too-consuming to rank all hashtags for every micro-video during evaluation, we followed the common strategy that randomly samples 100 hashtags that are not interacted by the micro-video, ranking the test hashtag among the 100 hashtags. Then each method outputs prediction scores for these 101 instances to rank them accordingly. To evaluate the ranking performance, we employed the widely used metric — Recall and NDCG. These metrics are on an interaction level, so the overall performance is the average result among all interactions. Larger values indicate the better performance. Recall measures whether the testing item is ranked in the top-k list, while NDCG accounts for the position of the hit by assigning a higher score to hit at top positions. The same settings are also applied for the hyper-parameter tuning on the validation set.

4.1.3. Baselines

To justify the effectiveness of our framework, we compared it with the following methods.

- **RSDAE** [34]. This is a textual modality-based hashtag recommendation algorithm, integrating the deep representation learning and relational learning under a probabilistic framework. We adopted the released implementation[9] and modified its evaluation scheme to adapt to our testing scenario.

[9] http://www.wanghao.in/publication.html.

- **Co-Attention** [32]. This is the state-of-the-art hashtag recommender system. It introduces a co-attention network, incorporating both the textual and visual information. We employed the implementation released by the authors[10] and modified its evaluation strategy.
- **TRUMANN** [15]. This is a tree-guided multi-task multi-modal learning approach designed to estimate the venue category for micro-videos. We modified the implementation released by the authors[11] by retaining the common space learning component and casting the representations of multiple modalities into a softmax classifier to perform the hashtag recommendation.
- **EASTERN** [18]. This is an end-to-end deep learning algorithm, equipped with three parallel RNNs to capture the sequential structures and a CNN to learn the sparse concept-level representations of micro-videos. We adopted the model design[12] and cascaded the feature vectors of multiple modalities into a softmax classifier to evaluate the hashtag recommendation performance.
- **LOGO-A, LOGO-M, and LOGO-L.** These are variants of our LOGO method by removing the attention component and employing the average pooling (LOGO-A), maximum pooling (LOGO-M), and last pooling (LOGO-L) as the integration scheme to represent the whole sequence. This is to demonstrate the effect of our proposed attention-based pooling.

4.1.4. Implementation and hyper-parameter setting

We implemented our method based on PyTorch.[13] To initialize the embedding layer and the hidden layers, we randomly set their parameters with a Gaussian distribution (a mean of 0 and a standard deviation of 0.1). We used the Adam optimizer for all gradient-based methods, where the mini-batch size and learning rate were searched in $[128, 256, 512, 1024]$ and $[0.001, 0.005, 0.01, 0.05, 0.1]$, respectively. In addition, the hidden units in visual, acoustic, and textual LSTMs are set as 500, 300, and 80, respectively, and the dimension of the common space is set as 150, consistent with the settings in [18]. For hidden layers in the three mapping functions (Eq. (4)) and the customized NCF (Eq. (7)), we empirically set the size of the first hidden layer same as the input embedding size, and employed three layers of a tower structure and ReLU activation function. Parameters of both our method and baselines are carefully tuned on the validation set to report their best performance. We repeated each setting for 5 times and reported the average results. Besides, we further calculated the standard deviation on both Recall and NDCG based on the 5 times experiments.

4.2. Overall performance comparison (RQ1)

Experimental results of various methods are summarized in Table 1. We have the following observations: (1) Our LOGO solution achieves the best performance on both Recall and NDCG, significantly outperforming other state-of-the-art methods.[14] This mainly because LOGO considers the sequential structure learning and multi-modal fusion. (2) The performance of LOGO is superior to that of LOGO-A, LOGO-M, and LOGO-L. This demonstrates the effectiveness of the attention mechanism, distinguishing the importance of sequential units in each modality. (3) The visual and textual modalities-based algorithm Co-Attention beats the textual modality-based algorithm RSDAE by a great margin. The visual modality is highly related to the hashtag recommendation task, but it is ignored in RSDAE. (4) Although the visual, acoustic, and textual modalities are incorporated in TRUMANN and EASTERN, their performance is still worse than that of LOGO and Co-Attention. TRUMANN and EASTERN are proposed for multi-modal fusion, while LOGO and Co-Attention are specifically designed for the hashtag recommendation task. It well validates the necessity of considering the specialities of hashtag recommendation, such as the interaction learning and attention mechanism.

4.3. Impact of popularity (RQ2)

To investigate how our proposed approach and other baselines behave for popular, relative popular, and unpopular hashtags, we further performed a micro-scope study. We sorted the hashtags according to the number of occurrences in descending order and divided them into three disjoint sets with the ratio of 7:2:1. In particular, if a hashtag occurs more than 27 times, the hashtag is classified as a popular one; if the number of occurrences of a hashtag locates in the range of $[10, 27]$, the hashtag is regarded as a relative popular one; otherwise a hashtag is defined as an unpopular one. Maintaining the same trained models before, Table 2 reveals the experimental results of various methods with respect to popular, relative popular, and unpopular hashtags. In addition, to avoid the possibility of trained models influenced by different types of hashtags, we divided the original dataset into three disjoint datasets with different types of hashtags and trained models on these datasets individually. Table 3 shows the experimental results of various methods on the datasets with popular hashtags, relative popular hashtags, and unpopular hashtags, respectively. Jointly observing Tables 2 and 3, we have the following observations: (1) The performance of LOGO is consistently and significantly superior to that of baselines for popular, relative popular, and unpopular hashtags. It well illustrates the superiority of our proposed framework. (2) The recommendation performance of popular hashtags surpasses that of relative popular hashtags and further beats unpopular hashtags. If a hashtag occurs more frequently in our training set, the representation of the hashtag is better learnt. That is why the performance of popular hashtags is superior to that of relative popular hashtags and unpopular hashtags. (3) Experimental results adequately evaluate the performance of various methods for different types of hashtag with two different dataset divisions. The conclusions are consistent in Tables 2 and 3, which demonstrates the robustness and effectiveness of our proposed solution.

4.4. Parameter tuning and sensitivity (RQ3)

To demonstrate the robustness and effectiveness of our LOGO framework, we investigated the sensibility of several factors, namely, the number of negative samples, the dropout ratio ρ, and the tradeoff parameter λ.

Impact of Negative Samples: The impact of negative sampling for LOGO is revealed in Fig. 5(a). We have the following observations: (1) One negative sample for each positive instance is not optimal for the final performance; and the more negative samples are sampled, the more benefits it brings. Compared with the traditional pairwise sampling method which selects only one negative sample to pair with each positive sample, such as BPR [47], LOGO shows the advantage of flexible negative samples. (2) With more negative samples selected, the performance of LOGO becomes stable and reaches its optimal result around 6. Therefore, we made 6 as the default negative sampling ratio for LOGO.

Impact of Dropout: To prevent LOGO from overfitting, we employed the dropout strategy to improve the regularization

[10] http://jkx.fudan.edu.cn/~qzhang/publications.html.

[11] http://acmmm16.wixsite.com/mm16.

[12] https://acmmm17.wixsite.com/eastern.

[13] http://www.pytorch.org.

[14] All the p-values between our model and each baseline are much smaller than 0.05, which indicates that the improvements are statistically significant.

Table 1
Overall performance comparison among various methods.

Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.4217	7.43e-02	0.3330	5.83e-02	0.5036	8.85e-02	0.3592	6.30e-02
Co-Attention	0.5282	2.11e-02	0.4088	2.04e-02	0.6441	1.83e-02	0.4588	1.32e-02
TRUMANN	0.4853	4.25e-02	0.3837	3.29e-02	0.6044	3.81e-02	0.4225	3.13e-02
EASTERN	0.5253	2.25e-02	0.4052	2.22e-02	0.6366	2.20e-02	0.4412	2.20e-02
LOGO-A	0.5474	1.12e-02	0.4334	8.15e-03	0.6644	8.25e-03	0.4693	8.05e-03
LOGO-M	0.5299	2.02e-02	0.4121	1.87e-02	0.6416	1.95e-02	0.4482	1.85e-02
LOGO-L	0.5491	1.07e-02	0.4197	1.49e-02	0.6588	1.10e-02	0.4647	1.03e-02
LOGO	**0.5705**	–	**0.4497**	–	**0.6809**	–	**0.4854**	–

Table 2
Performance comparison among various methods w.r.t. popular, relative popular, and unpopular hashtags with the same training set and different testing sets.

Popular hashtags

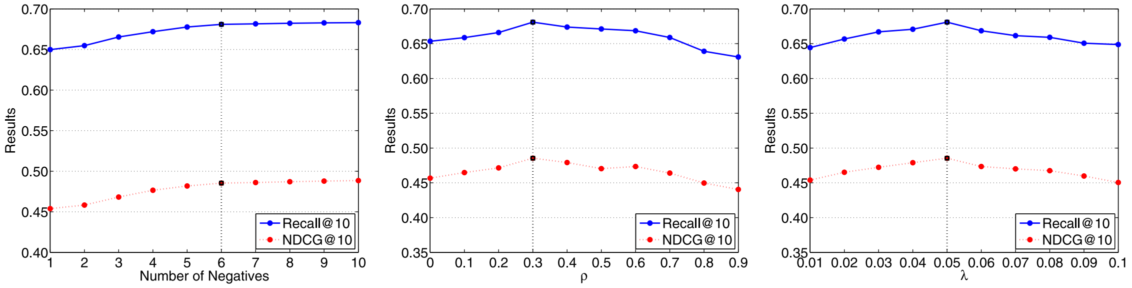
Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.4421	7.82e-02	0.3461	5.62e-02	0.5277	8.15e-02	0.3764	6.13e-02
Co-Attention	0.5517	2.34e-02	0.4264	1.61e-02	0.6725	9.24e-03	0.4788	1.01e-02
TRUMANN	0.5503	2.41e-02	0.4221	1.82e-02	0.6691	1.09e-02	0.4701	1.45e-02
EASTERN	0.5253	3.66e-02	0.4052	2.67e-02	0.6366	2.71e-02	0.4412	2.89e-02
LOGO	**0.5988**	–	**0.4588**	–	**0.6910**	–	**0.4992**	–

Relative popular hashtags

Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.4332	7.97e-02	0.3357	5.66e-02	0.5213	7.86e-02	0.3718	5.46e-02
Co-Attention	0.5450	2.38e-02	0.4165	1.62e-02	0.6624	8.20e-03	0.4720	4.68e-03
TRUMANN	0.5024	4.51e-02	0.3887	3.01e-02	0.6257	2.64e-02	0.4326	2.42e-02
EASTERN	0.5453	2.36e-02	0.4123	1.83e-02	0.6550	1.18e-02	0.4639	8.64e-03
LOGO	**0.5928**	–	**0.4491**	–	**0.6788**	–	**0.4812**	–

Unpopular hashtags

Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.3122	6.57e-02	0.2571	4.19e-02	0.4109	6.49e-02	0.2909	5.50e-02
Co-Attention	0.4122	1.58e-02	0.3190	1.10e-02	0.5244	8.30e-03	0.3902	5.50e-03
TRUMANN	0.3774	3.31e-02	0.2937	2.36e-02	0.4831	2.88e-02	0.3566	2.21e-02
EASTERN	0.4122	1.58e-02	0.3009	2.00e-02	0.5244	8.30e-03	0.3812	9.93e-03
LOGO	**0.4439**	–	**0.3411**	–	**0.5410**	–	**0.4011**	–



(a) Negative Samples (b) Dropout Ratio (c) Tradeoff Parameter

Fig. 5. Performance of LOGO w.r.t. the number of negatives, the dropout ratio ρ, and the tradeoff parameter λ on both Recall@10 and NDCG@10.

of our deep model. In particular, we randomly dropped ρ of neurons on the pooling layers and hidden layers, whereinto ρ is the dropout ratio. Fig. 5(b) exhibits the performance of LOGO w.r.t. the dropout ratio ρ. We have the following observations: (1) When the dropout ratio equals to 0, LOGO performs poorly due to the overfitting. (2) The optimal setting for the dropout ratio locates at 0.3. When the dropout ratio exceeds the optimal setting, the performance of LOGO decreases dramatically because of the insufficient information.

Impact of Tradeoff Parameter: λ balances the modality similarity influence of micro-videos and hashtags. The parameter tuning result of λ is illustrated in Fig. 5(c). As can be seen, both Recall@10 and NDCG@10 increase first and then decrease, and reach their maximum values when $\lambda = 0.05$. With the increase of λ, our model puts more emphasis on the modality similarity influence of hashtags. However, over emphasis on the influence is unreasonable. This is why the values of Recall@10 and NDCG@10 decrease after $\lambda = 0.05$. Therefore, we chose $\lambda = 0.05$ as our default value.

Table 3
Performance comparison among various methods w.r.t. popular, relative popular, and unpopular hashtags with different training sets and different testing sets.

Dataset of popular hashtags								
Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.4536	8.13*e*-02	0.3640	5.74*e*-02	0.5488	8.06*e*-02	0.3796	6.50*e*-02
Co-Attention	0.5706	2.28*e*-02	0.4415	1.87*e*-02	0.6996	5.35*e*-03	0.4848	1.25*e*-02
TRUMANN	0.5241	4.60*e*-02	0.4146	3.21*e*-02	0.6522	2.89*e*-02	0.4563	2.67*e*-02
EASTERN	0.5674	2.44*e*-02	0.4376	2.06*e*-02	0.6878	1.12*e*-02	0.4764	1.67*e*-02
LOGO	**0.6164**	–	**0.4791**	–	**0.7102**	–	**0.5099**	–
Dataset of relative popular hashtags								
Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.4172	7.48*e*-02	0.3348	5.28*e*-02	0.5048	7.41*e*-02	0.3492	5.98*e*-02
Co-Attention	0.5228	2.20*e*-02	0.4025	1.90*e*-02	0.6327	1.02*e*-02	0.4383	1.53*e*-02
TRUMANN	0.4829	4.19*e*-02	0.3814	2.95*e*-02	0.6004	2.63*e*-02	0.4198	2.45*e*-02
EASTERN	0.5249	2.09*e*-02	0.4068	1.68*e*-02	0.6436	4.92*e*-03	0.4460	1.15*e*-02
LOGO	**0.5670**	–	**0.4407**	–	**0.6533**	–	**0.4691**	–
Dataset of unpopular hashtags								
Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
RSDAE	0.3059	6.44*e*-02	0.2519	4.10*e*-02	0.4026	6.40*e*-02	0.2850	5.63*e*-02
Co-Attention	0.4028	1.60*e*-02	0.2948	1.96*e*-02	0.5112	9.78*e*-03	0.3735	1.21*e*-02
TRUMANN	0.3698	3.25*e*-02	0.2878	2.31*e*-02	0.4734	2.86*e*-02	0.3494	2.41*e*-02
EASTERN	0.4056	1.46*e*-02	0.3122	1.09*e*-02	0.5139	8.45*e*-03	0.3824	7.71*e*-03
LOGO	**0.4350**	–	**0.3342**	–	**0.5308**	–	**0.3978**	–

Table 4
Performance comparison of various modality combinations.

Methods	K=5				K=10			
	Recall	Deviation	NDCG	Deviation	Recall	Deviation	NDCG	Deviation
Visual	0.4758	4.72*e*-02	0.3827	3.34*e*-02	0.5452	6.77*e*-02	0.3916	4.68*e*-02
Acoustic	0.4616	5.43*e*-02	0.3684	4.05*e*-02	0.5233	7.87*e*-02	0.3746	5.53*e*-02
Textual	0.4609	5.47*e*-02	0.3675	4.10*e*-02	0.5234	7.86*e*-02	0.3742	5.55*e*-02
Acoustic+Textual	0.5163	2.70*e*-02	0.4020	2.37*e*-02	0.6072	3.67*e*-02	0.4178	3.37*e*-02
Visual+Textual	0.5541	8.20*e*-03	0.4341	7.81*e*-03	0.6590	1.09*e*-02	0.4707	7.36*e*-03
Visual+Acoustic	0.5588	5.89*e*-03	0.4412	4.34*e*-03	0.6649	8.00*e*-03	0.4761	4.73*e*-03
All	**0.5705**	–	**0.4497**	–	**0.6809**	–	**0.4854**	–

4.5. *Evaluation on modality combination (RQ4)*

To demonstrate the effectiveness of our proposed multi-modal fusion, we performed the study on various modality combinations. The performance of different modality combinations is shown in Table 4. We have the following observations: (1) In terms of single modality comparison, the performance of *Visual* is significantly superior to that of *Acoustic* and *Textual*. This is mainly because the visual modality is the primary information carrier of micro-videos and thus dominates the hashtag recommendation performance. Meanwhile, it implies that the visual features extracted from the AlexNet [38] are capable of capturing high-quality hashtag-related information. (2) The performance of *Acoustic* and *Textual* is rather close, and the performance of *Visual+Acoustic* is slightly better than that of *Visual+Textual*. It demonstrates that the acoustic and textual features extracted from acoustic and textual modalities are rather close in performing the hashtag recommendation. (3) It is obviously observed that the more modalities are incorporated, the better performance we achieved. This verifies that there are consistent rather than conflicting relationships among multiple modalities. (4) The method of *All* achieves the best performance among various modality combinations. This validates the effectiveness of our LOGO method, more specifically, the efficiency in aggregating multiple modalities.

5. Conclusions and future work

In this paper, we present an end-to-end solution, LOGO, to solve the hashtag recommendation issue in the micro-video context by utilizing multiple modalities (i.e., visual, acoustic, and textual modalities). Under this framework, two key issues are well-addressed, namely, sequential structure modeling and multi-modal fusion. In particular, the features extracted from multiple modalities are fed into an attentive sequential structure learning component to capture the sequential and attentive features simultaneously. Meanwhile, a multi-view representation learning approach is proposed to fuse the multiple modalities by projecting their vector representations into a common space under the restriction of the modality similarity. Furthermore, we feed the projections of multiple modalities in the common space and the embeddings of hashtags into our customized NCF. The experimental results demonstrate that LOGO achieves the state-of-the-art performance for the hashtag recommendation; further micro-level analyses demonstrate how LOGO is sensitive to hyper-parameters and how LOGO performs with different modality combinations.

In future, we plan to extend our work in the following three directions. First, we will study how to employ the semantic information of hashtags instead of regarding them as symbols. Hashtags generally represent current trending topics or events. Therefore, we plan to transfer the external knowledge of hashtags to strengthen the hashtag recommendation performance. Second, we are interested in discovering the complex relatedness among the hashtags, such as the tree structure and graph structure. By leveraging the predefined structure to regularize the relatedness among the hashtags, the hashtag recommendation performance may be further enhanced. Third, we plan to perform the hashtag recommendation for micro-videos in an online manner. The content of micro-videos changes over time, so do the hashtags. It would be helpful to utilize users' reviews to capture the dynamic changes of both micro-videos and hashtags.

CRediT authorship contribution statement

Da Cao: Conceptualization, Methodology. **Lianhai Miao:** Formal analysis, Validation. **Huigui Rong:** Resources, Data curation. **Zheng Qin:** Supervision, Funding acquisition. **Liqiang Nie:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are highly grateful to the anonymous reviewers for their careful reading and insightful comments. This work is supported by the National Key Research and Development Project of China (No. 2018YFB0704000 and No. 2017YFB0902904), the National Natural Science Foundation of China (No. 61802121, No. 61672221, and No. 61772191), the Natural Science Foundation of Hunan Province (No. 2019JJ50057, No. 2020JJ4008, and No. 2018JJ3259), the Science and Technology Key Projects of Hunan Province (No. 2015TP1004, No. 2018TP1009, No. 2019GK2082, No. 2018TP2023, and No. 2018TP3001), the Transportation Science and Technology Project of Hunan Province (No. 201819, kq1804008), and the Fundamental Research Funds for the Central Universities.

References

[1] D. Kowald, S.C. Pujari, E. Lex, Temporal effects on hashtag reuse in twitter: A cognitive-inspired hashtag recommendation approach, in: Proceedings of the International Conference on World Wide Web, ACM, 2017, pp. 1401–1410.
[2] J. Zhang, Y. Yang, Q. Tian, L. Zhuo, X. Liu, Personalized social image recommendation method based on user-image-tag model, IEEE Trans. Multimed. 19 (11) (2017) 2439–2449.
[3] B. Shi, G. Poghosyan, G. Ifrim, N. Hurley, Hashtagger+: Efficient high-coverage social tagging of streaming news, IEEE Trans. Knowl. Data Eng. 30 (1) (2018) 43–58.
[4] S. Hochreiter, J. Schmidhuber, Lstm can solve hard long time lag problems, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, 1997, pp. 473–479.
[5] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, 2014, pp. 3104–3112.
[6] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–15.
[7] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.
[8] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133.
[9] Y. Fang, H. Zhang, Y. Ren, Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing, Knowl.-Based Syst. 171 (2019) 69–80.
[10] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross-modal retrieval, Knowl.-Based Syst. 180 (2019) 38–50.
[11] F. Huang, X. Zhang, J. Xu, C. Li, Z. Li, Network embedding by fusing multimodal contents and links, Knowl.-Based Syst. 171 (2019) 44–55.
[12] F. Huang, X. Zhang, Z. Li, Z. Zhao, Y. He, From content to links: Social image embedding with deep multimodal model, Knowl.-Based Syst. 160 (2018) 251–264.
[13] N. Takahashi, M. Gygli, L.V. Gool, Aenet: Learning deep audio features for video analysis, IEEE Trans. Multimed. 20 (3) (2018) 513–524.
[14] F. Wang, H. Nagano, K. Kashino, T. Igarashi, Visualizing video sounds with sound word animation to enrich user experience, IEEE Trans. Multimed. 19 (2) (2017) 418–429.
[15] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, T.S. Chua, Shorter-is-better: Venue category estimation from micro-video, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2016, pp. 1415–1424.
[16] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, T.S. Chua, Micro tells macro: Predicting the popularity of micro-videos via a transductive model, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2016, pp. 898–907.
[17] J. Wu, Y. Zhou, Z. Zhu, Z. Zhu, Modeling dynamics of online video popularity, IEEE Trans. Multimed. 18 (9) (2016) 1882–1895.
[18] M. Liu, L. Nie, M. Wang, B. Chen, Towards micro-video understanding by joint sequential-sparse modeling, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2017, pp. 970–978.
[19] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, Q. Tian, Enhancing micro-video understanding by harnessing external sounds, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2017, pp. 1192–1200.
[20] B. Samanta, A. De, N. Ganguly, Strm: A sister tweet reinforcement process for modeling hashtag popularity, in: Proceedings of the IEEE International Conference on Computer Communications, IEEE, 2017, pp. 1–9.
[21] Y. Gao, J. Sang, T. Ren, C. Xu, Hashtag-centric immersive search on social media, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2017, pp. 1924–1932.
[22] C. Xing, Y. Wang, J. Liu, Y. Huang, W.-Y. Ma, Hashtag-based sub-event discovery using mutually generative lda in twitter, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 2666–2672.
[23] Mingsong Mao, Jie Lu, Guangquan Zhang, Jinlong Zhang, Multirelational social recommendations via multigraph ranking, IEEE Trans. Cybern. 47 (12) (2016) 4049–4061.
[24] Qian Zhang, Dianshuang Wu, Jie Lu, Feng Liu, Guangquan Zhang, A cross-domain recommender system with consistent information transfer, Decis. Support Syst. 104 (2017) 49–63.
[25] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, Jun Ma, Neural attentive session-based recommendation, in: Proceedings of the ACM Conference on Information and Knowledge Management, ACM, 2017, pp. 1419–1428.
[26] Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, Chi Xu, Proceedings of the ACM Conference on Recommender Systems, ACM, 2018, pp. 297–305.
[27] Z. Ding, X. Qiu, Q. Zhang, X. Huang, Learning topical translation model for microblog hashtag suggestion, in: Proceedings of International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2078–2084.
[28] H. Wang, B. Chen, W.J. Li, Collaborative topic regression with social regularization for tag recommendation, in: Proceedings of International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2719–2725.
[29] Y. Gong, Q. Zhang, X. Huang, Hashtag recommendation for multimodal microblog posts, Neurocomputing 272 (2018) 170–177.
[30] Y. Gong, Q. Zhang, X. Huang, Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL, 2015, pp. 401–410..
[31] Y. Gong, Q. Zhang, Hashtag recommendation using attention-based convolutional neural network, in: Proceedings of International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 2782–2788.
[32] Q. Zhang, J. Wang, H. Huang, X. Huang, Y. Gong, Hashtag recommendation for multimodal microblog using co-attention network, in: Proceedings of International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 3420–3426.
[33] J. Li, H. Xu, X. He, J. Deng, X. Sun, Tweet modeling with lstm recurrent neural networks for hashtag recommendation, in: Proceedings of the Internation Joint Conference on Neural Networks, IEEE, 2016, pp. 1570–1577.
[34] H. Wang, X. Shi, D.Y. Yeung, Relational stacked denoising autoencoder for tag recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2015, pp. 3052–3058.
[35] Y. Li, T. Liu, J. Jiang, L. Zhang, Hashtag recommendation with topical attention-based LSTM, in: Proceedings of the International Conference on Computational Linguistics, ACM, 2016, pp. 3019–3029.
[36] L. Huang, B. Luo, Tag refinement of micro-videos by learning from multiple data sources, Multimedia Tools Appl. 76 (19) (2017) 20341–20358.

[37] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, L. Nie, Personalized hashtag recommendation for micro-videos, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2019, pp. 1446–1454.

[38] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, 2012, pp. 1097–1105.

[39] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, 2013, pp. 3111–3119.

[40] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, T.-S. Chua, Attentional factorization machines: learning the weight of feature interactions via attention networks, in: Proceedings of International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 3119–3125.

[41] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the International Conference on World Wide Web, ACM, 2017, pp. 173–182.

[42] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (8) (2009).

[43] X. He, T.-S. Chua, Neural factorization machines for sparse predictive analytics, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 355–364.

[44] P. Covington, J. Adams, E. Sargin, Deep neural networks for youtube recommendations, in: Proceedings of the ACM Conference on Recommender Systems, ACM, 2016, pp. 191–198.

[45] X. Wang, X. He, F. Feng, L. Nie, T.S. Chua, TEM: Tree-enhanced embedding model for explainable recommendation, in: Proceedings of the International Conference on World Wide Web, ACM, 2018, pp. 1543–1552.

[46] D. Cao, X. He, L. Miao, Y. An, C. Yang, R. Hong, Attentive group recommendation, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2018.

[47] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the International Conference on Uncertainty in Artificial Intelligence, AUAI, 2009, pp. 452–461.