Long-tail Hashtag Recommendation for Micro-videos with Graph Convolutional Network

Mengmeng Li
Shandong University
doubleflmm@gmail.com

Tian Gan*
Shandong University
gantian@sdu.edu.cn

Meng Liu
Shandong University
mengliu.sdu@gmail.com

Zhiyong Cheng
Qilu University of Technology
(Shandong Academy of Sciences)
jason.zy.cheng@gmail.com

Jianhua Yin
Shandong University
jhyin@sdu.edu.cn

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

ABSTRACT

Hashtags, a user provides to a micro-video, are the ones which can well describe the semantics of the micro-video's content in his/her mind. At the same time, hashtags have been widely used to facilitate various micro-video retrieval scenarios (*e.g.,* search, browse, and categorization). Despite their importance, numerous micro-videos lack hashtags or contain inaccurate or incomplete hashtags. In light of this, hashtag recommendation, which suggests a list of hashtags to a user when he/she wants to annotate a post, becomes a crucial research problem. However, little attention has been paid to micro-video hashtag recommendation, mainly due to the following three reasons: 1) lack of benchmark dataset; 2) the temporal and multi-modality characteristics of micro-videos; and 3) hashtag sparsity and long-tail distributions. In this paper, we recommend hashtags for micro-videos by presenting a novel multi-view representation interactive embedding model with graph-based information propagation. It is capable of boosting the performance of micro-videos hashtag recommendation by jointly considering the sequential feature learning, the video-user-hashtag interaction, and the hashtag correlations. Extensive experiments on a constructed dataset demonstrate our proposed method outperforms state-of-the-art baselines. As a side research contribution, we have released our dataset and codes to facilitate the research in this community.

KEYWORDS

Micro-videos, Hashtag Recommendation, Long-tail

ACM Reference Format:
Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. 2019. Long-tail Hashtag Recommendation for Micro-videos with Graph Convolutional Network. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357912

*Tian Gan and Liqiang Nie are corresponding authors.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357912

1 INTRODUCTION

Micro-videos, as a new trend in user-generated content, have been widely spread on various social platforms, such as Instagram and Snapchat [28, 40, 48]. The micro-videos on these social platforms are usually associated with hashtags, which are commonly used to summarize the content of micro-videos and attract the attention of followers [31]. Taking the popular social platform Instagram as an example, the hashtags are prefixed with the symbol "#" to mark keywords or key topics of a post. The hashtags have been proved to be useful in many applications, including microblog retrieval [8], event analysis [43], and sentiment analysis [38]. Furthermore, the tagging service can benefit the stakeholders of micro-video ecosystems. For users, the hashtags facilitate the search and location of their desired micro-videos. For the post-sharers, concise and concrete hashtags can increase the probability of their micro-videos to be discovered. For platforms, the hashtags can make the management of micro-videos (e.g., categorization) more convenient. Unfortunately, many users do not provide hashtags to their posts. To facilitate the usage of hashtags, hashtag recommendation has become an important research topic with considerable attention in recent years.

Several models have been adopted for hashtag recommendation, such as collaborative filtering [21], generative models [7, 14], and deep neural networks [13, 27, 37, 46]. Although some progress has been achieved so far, they mainly focus on the hashtag recommendation for microblogs or social images. However, recommending hashtags for micro-videos is non-trivial due to the following challenges: **1) Long-tail distribution.** The hashtag distribution is heavily skewed towards a few frequent hashtags with a long-tail consisting of less frequent tags [37]. Current studies note that many tags from the long-tail are "misspelled" or "meaningless" words [41], yet we believe that there are some meaningful hashtags within the long-tail which have been overlooked. That is, how to create correlations between the frequent hashtags and their "related" long-tail hashtags to enhance the representation of them is untapped. **2) Multimodal Sequence Modeling.** Micro-videos consist of visual, acoustic, and textual modalities, which are encoded together with sequential structure (*i.e.,* a set of ordered image frames, a list of audio clips with successive amplitude of wave, and a series of semantically and syntactically correlated words). Different streams in a micro-video imply different temporal dynamics and should therefore be modeled separately. For example, a micro-video contain the same objects over the time span of itself, while the actions and audio may change at intervals. Meanwhile, different modalities depict the intrinsic content of micro-videos consistently and complementarily from different views. Therefore, how to capture sequential and multi-modality features is a considerable problem. To address the aforementioned problems, we propose a multi-view interactive embedding personalized hashtag recommendation model with graph-guided information propagation.

The overview of our proposed method is illustrated in Figure 1. We first constructed a graph to explore hashtag correlations with external knowledge, and then leveraged existing structural knowledge to derive proper dependencies between frequent hashtags and long-tail hashtags. The propagation of such hashtag relation information was then used to modify the representation of the initial hashtag representation. Afterwards, we utilized three parallel Long Short-Term Memory Networks (LSTMs) to model the sequential features for units in each modality and the outputs of the three LSTMs were projected into a common space. Finally, we employed an interactive embedding network to predict the interactions among hashtags, micro-videos, and users.

As far as we know, this is the first work to recommend hashtags for micro-videos. By conducting experiments on our constructed real-world dataset, our proposed approach has demonstrated significant gains as compared with other hashtag recommendation approaches. The main contributions are summarized as follows:

(1) We proposed a joint framework that incorporates micro-videos, hashtags, and users to recommend hashtags. By projecting their representations into the same space and exploiting their interactions explicitly, our proposed model achieves better results on this task.
(2) We introduced a novel method to successively address the hashtag long-tail phenomenon by constructing a hashtag graph with external knowledge and integrating a propagation mechanism to exploit hashtag correlations.
(3) We built a large-scale micro-video dataset with a large hashtag vocabulary and released it to facilitate the research community[1].

2 RELATED WORK

Our study is related to prior studies on 1) hashtag recommendation, and 2) long tail recommendation.

2.1 Hashtag Recommendation

Hashtags are widely used in various scenarios, such as popularity prediction [35, 44], immersive search [12], and enterprise applications [29]. Generally speaking, prior efforts in hashtag recommendation can be divided into two categories based on their associated data: microblogs (e.g., Twitter and Sina-Weibo), and social images (e.g., Flicker and Facebook).

Hashtag recommendation for microblogs has been proposed from different perspectives, including collaborative filtering [21], generative models [7, 14], and neural network-based models [26, 46]. Collaborative filtering is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users. Kywe *et al.* [21] proposed a collaborative filtering method to recommend hashtags by combining hashtags from similar tweets and the ones from similar users. Generative models exploit the hashtags by modeling the hashtag generating process via the probability theory. Ding *et al.* [7] modeled the hashtag recommendation task as a translation process through extending the translation based method and introducing a topic-specific translation model to process the various meanings of words in different topics. Gong *et al.* [14] proposed that different types of hashtags follow different distributions and then they incorporated these hashtags into the topical translation model for hashtag recommendation task. Different from generative models, neural network-based models explore the hashtag recommendation task by utilizing the techniques on deep neural networks, such as attention mechanism and sequential learning. For example, a co-attention network is proposed in [46] to recommend hashtags for multimodal tweets by incorporating textual and visual information; an attention-based LSTM in [26] incorporates topic modeling into the LSTM architecture through an attention mechanism.

Apart from recommending hashtags for microblogs, many efforts have been done on recommendation for social images. Motivated by the fact that data labels hashtags are inherently related, Wang *et al.* [39] presented a joint framework that predicts class labels and hashtags for social media posts simultaneously. Rawat *et al.* [33] proposed a context-aware model to integrate context information with image content for multi-label hashtag prediction. Veit *et al.* [37] and Denton *et al.* [6] incorporated images, hashtags, and users into a three-way tensor model to model the interaction among image features, hashtag embedding, and user embedding.

2.2 Long-tail Recommendation

It is well-known that the frequency of objects occurring in natural scenes follows a long-tail distribution [34]. Long-tails complicate the analysis because rare cases from the tail still collectively make up a significant portion of the data and hence cannot be ignored [47]. In recent years, the long-tail problem has been widely investigated in recommendation systems and multi-label recognition.

Park [32] proposed an adaptive clustering method, in which the recommendations for long-tail items are based on the ratings in more intensively clustered groups and the frequent items are based on the ratings of individual items. Kordumova *et al.* [20] investigated what social tags constitute the long tail and how they perform on two multimedia retrieval scenarios, tag relevance, and detector learning. By augmenting the rare tags with simple semantics, the performance of tag relevance and detector learning improves considerably. Considering that accuracy is insufficient in assessing the quality, some studies [15, 36] exploit recommendations by considering other criteria in addition to accuracy. Shi [36] proposed a graph-based recommendation to effectively trade off between accuracy and long-tail. Another study [18] recommends lists ranked according to five dimensions which are accuracy, balance (e.g., the distribution of recommendations among all items), item coverage, quantity, and quality of long-tail item recommendation.

[1]https://anon425.wixsite.com/v2ht

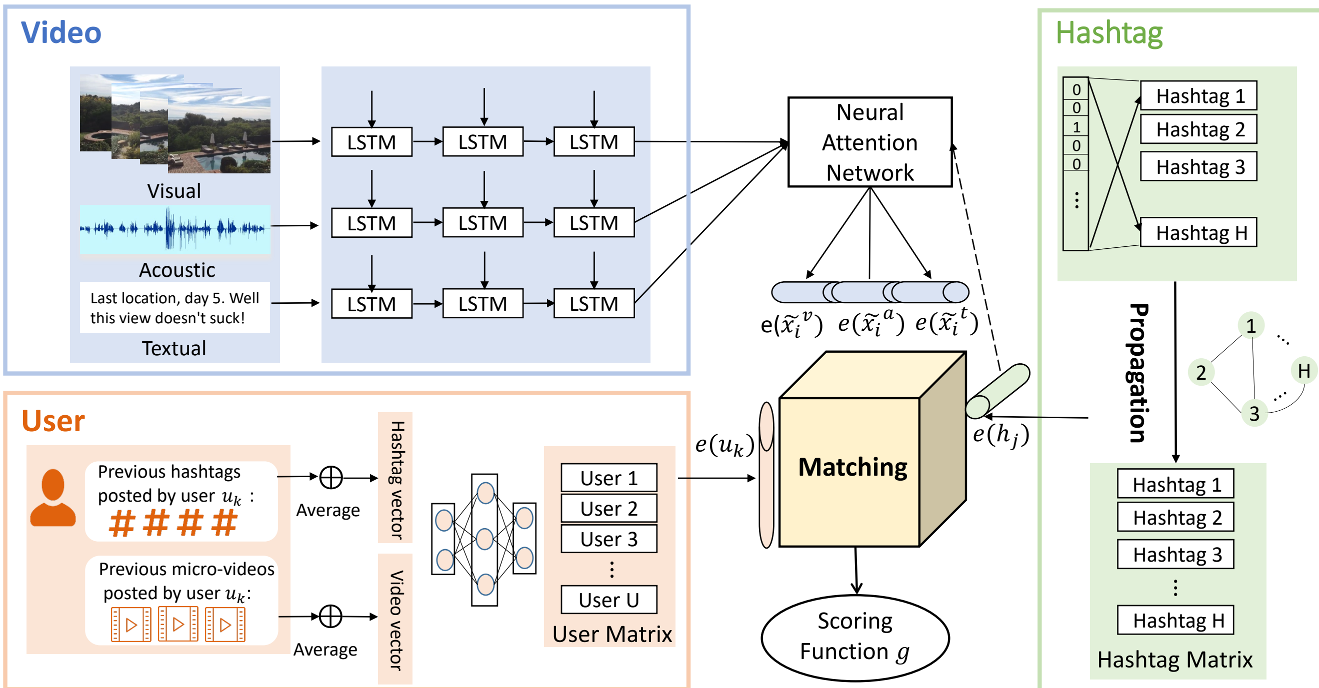


Figure 1: Overview of the proposed model for hashtag recommendation.

Hamedani *et al.* [15] proposed an approach in which the recommendation list is optimized based on three objectives: increasing the accuracy, personalizing the diversity, and reducing the popularity of the recommended items to serve the purpose.

The long-tail problem in multi-label recognition is also a challenge. A straightforward way for multi-label recognition is to train independent binary classifiers for each class/label. However, this method does not consider the relationship among labels. To enhance the representation of long-tail labels, many researchers attempted to use label co-occurrence and semantic relations between labels to capture label dependencies with the graph. Li *et al.* [24] created a tree-structured graph in the label space by using the maximum spanning tree algorithm. Li *et al.* [25] produced image-dependent conditional label structures on the basis of the graphical Lasso framework. Lee *et al.* [22] incorporated knowledge graphs for describing the relationships among multiple labels.

3 OUR PROPOSED FRAMEWORK

3.1 Overview

In this work, we propose an interactive model that incorporates hashtags, micro-videos, and users simultaneously for micro-videos hashtag recommendation. Formally, we assume a set of micro-videos $\mathcal{V} = \{v_1, v_2, ..., v_{|V|}\}$, a vocabulary of hashtags $\mathcal{H} = \{h_1, h_2, ..., h_{|H|}\}$, and a set of users $\mathcal{U} = \{u_1, u_2, ..., u_{|U|}\}$, where $|\cdot|$ denotes the cardinality of a set. We further define a triplet $\tau = (v_i, h_j, u_k) \in \mathcal{Q}^+$ as a valid interaction if user u_k has added hashtag h_j for its posted micro-video vi_i, otherwise, $\tau \in \mathcal{Q}^-$. Each micro-video is associated with one or more hashtags posted by a unique user. The goal of our model is to predict a score $g(\tau)$ for each triplet, such that for any triplet pair $\tau^+ \in \mathcal{Q}^+$ and $\tau^- \in \mathcal{Q}^-$, $g(\tau^+) > g(\tau^-)$. The attention mechanism is introduced in the multimodal feature learning to filter out noises and find information that is most relevant to the corresponding hashtags.

3.2 Hashtag Embedding

The frequency of hashtags for micro-videos has an imbalanced distribution. For example, the frequent hashtags appear thousands of times (*e.g.*, #love, #fitness, and #music), while the rare ones only appear a few times (*e.g.,* #warsaw, #kennedy paige, and #light drip). The uneven distribution means that few common hashtags will dominate any error measure, and make it hard to predict rare hashtags at the long-tail [6]. In this work, we address the long-tail distribution issue by hashtag embedding propagation with external knowledge. Specifically, we first construct a graph by exploring hashtag correlations. Then, we introduce a propagation mechanism using the constructed graph. The core idea is that the frequent hashtags are capable of sharing knowledge to their "related" long-tail hashtags. Formally, let $\boldsymbol{e}(h_j) \in \mathbb{R}^{d_D}$ represent the initial embedding of the j-th hashtag, where d_D denotes the dimension of the hashtag embedding. After the propagation, $\boldsymbol{e}(h_j)$ will be translated into new representation $\boldsymbol{e}'(h_j)$ with shared knowledge encoded. Figure 2 illustrates the propagation mechanism.

3.2.1 *Graph Convolutional Network.*

Graph Convolutional Network (GCN) is introduced in [19] for the task of semi-supervised classification. In their work, GCN is used to generate node embedding in the graph based on local neighborhoods. Unlike standard

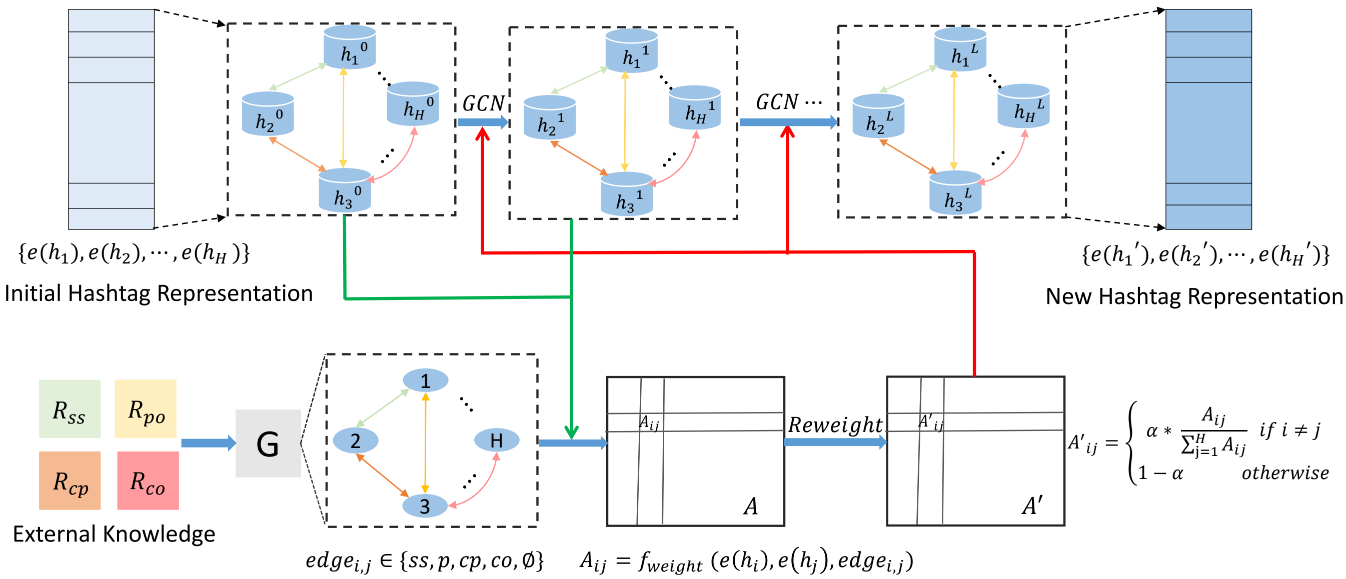


Figure 2: Illustration of the propagation mechanism in hashtag embedding module. A Graph G is built over the hashtags representations, where each node denotes a hashtag. Stacked GCNs are learned over the graph to map initial hashtag representations $\{e(h_1), e(h_2), ..., e(h_H)\}$ to new representations $\{e'(h_1), e'(h_2), ..., e'(h_H)\}$ with knowledge encoded.

convolutions that operate on local Euclidean structures in an image, the goal of GCN is to learn a function $f(\cdot,\cdot)$ on a Graph G, which takes feature descriptions $\boldsymbol{Y}^l \in \mathbb{R}^{b \times d_r^l}$ and the corresponding correlation matrix $\boldsymbol{A} \in \mathbb{R}^{b \times b}$ as inputs (where l denotes the layer of GCN, b represents the number of nodes, and d_r^l denotes the dimension of the l-th layer node features), and updates the node features as $\boldsymbol{Y}^{l+1} \in \mathbb{R}^{b \times d_r^{l+1}}$. The propagation rule for GCN layers can be written as a non-linear function by,

$$\boldsymbol{Y}^{l+1} = f(\boldsymbol{Y}^l, \boldsymbol{A}). \tag{1}$$

In particular, Kipf et al. [19] proposed a simple and well-behaved layer-wise propagation rule for neural network models, where $f(\cdot,\cdot)$ is represented as,

$$\boldsymbol{Y}^{l+1} = \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{Y}^l \boldsymbol{W}^l_{GCN}), \tag{2}$$

where $\boldsymbol{W}^l_{GCN} \in \mathbb{R}^{d_r^l \times d_r^{l+1}}$ is a layer-specific trainable weight matrix, $\tilde{\boldsymbol{A}} \in \mathbb{R}^{b \times b} = \boldsymbol{A} + \boldsymbol{I}_N$ is the adjacency matrix of the undirected graph G with added self-connections, $\boldsymbol{I}_N$ is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, and $\sigma(\cdot)$ denotes the activation function.

3.2.2 GCN for Hashtag Propagation. With the propagation rule defined in Equation (2), hashtags can aggregate information from their "related" neighbors and update their representations by stacking multiple GCN layers. However, a correlation matrix is required for the propagation.

In this work, we define the correlation matrix through a data-driven way. In particular, we first define four types of relations over the hashtag in the dataset, and then use these relations to build a graph for propagation. The four types of relations are as follows:

(1) *composition relation (cp)*, exists between unigram and n-gram. If the hashtag is composed of two or more words, there are connections between the hashtag and the words within it.
(2) *super-subordinate relation (ss)*, also called hyponymy, hypernomy or ISA relation, is defined in WordNet and can be extracted from it directly.
(3) *positive relation (po)*, exists among class labels. Label similarities are calculated by WUP similarity [42], followed by thresholding the soft similarities into positive relation.
(4) *co-occurrence relation (co)*, is defined by the co-occurrence of the hashtags. Two hashtags are defined as co-occurred if they appear in the same post.

The priorities of four relations are in the order we define them. That is, if a pair of hashtags contains multiple relations, we retain the one with the highest priority. The intuition behind these four relations is that the unigram hashtags are connected with their corresponding n-gram hashtag with the composition relation; the long-tail hashtags are augmented with semantically similar hashtags by the super-subordinate and positive relations; and the co-occurrence relation captures the weak connections among hashtags.

With the relations defined above, we construct four types of edges. Formally, G represents the graph, and $\{cp, ss, po, co\}$ are the types of edges in the graph. We denote G's correlation matrix as $\boldsymbol{A} \in \mathbb{R}^{H \times H}$, where H is the number of hashtags. We assign different weights on different types of edges. Specifically, given a pair of nodes j_1 and j_2, the propagation weight $A_{j_1 j_2}$ is determined by:

$$edge_{j_1,j_2} \in \{cp, ss, po, co, \phi\}, \tag{3}$$

$$A_{j_1 j_2} = f_{weight}(\boldsymbol{e}(h_{j_1}), \boldsymbol{e}(h_{j_2}), edge_{j_1,j_2}), \tag{4}$$

where $edge_{j_1,j_2}$ is the edge between the nodes j_1 and j_2, and function $f_{weight}(\cdot,\cdot)$ is used to compute the propagation weights, which is approximated by the neural networks.

A node would aggregate information from only relevant nodes that are defined in the graph to update its own hidden state vector. However, the aggregated representation of a node does not contain its own feature. Thus, we re-weight the correlation matrix $\boldsymbol{A}$, so that every node in the graph can combine its own prior representation. In particular, we adopted the re-weighted scheme in [2] that defines the re-weighted correlation matrix $\boldsymbol{A'}$ as,

$$A'_{j_1 j_2} = \begin{cases} \alpha \cdot \frac{A_{j_1 j_2}}{\sum_{j_2=1}^{H} A_{j_1 j_2}}, & if\ j_1 \neq j_2 \\ 1-\alpha, & otherwise \end{cases}, \tag{5}$$

where α determines the weights assigned to a node itself and other correlated nodes. By doing this, when $\alpha \longrightarrow 1$, the feature of a node itself will be ignored; when $\alpha \longrightarrow 0$, neighboring information will be ignored.

With the notation defined above, the propagation mechanism for hashtag representation is formulated as follows:

$$Y^0 = \{\boldsymbol{e}(h_1), \boldsymbol{e}(h_2), ..., \boldsymbol{e}(h_H)\}, \tag{6}$$

$$Y^{l+1} = h(\tilde{\boldsymbol{D}'}^{-\frac{1}{2}} \boldsymbol{A'} \tilde{\boldsymbol{D}'}^{-\frac{1}{2}} Y^l \boldsymbol{W}^l_{GCN}), \tag{7}$$

where $\{\boldsymbol{e}(h_1), \boldsymbol{e}(h_2), ..., \boldsymbol{e}(h_H)\}$ denotes the initial representation of hashtag. At last, we obtain the final hashtag representation by taking out the output of the last layer, *i.e.,* $\{\boldsymbol{e'}(h_1), \boldsymbol{e'}(h_2), ..., \boldsymbol{e'}(h_H)\}$.

3.3 Micro-video Embedding

Multi-view representation learning is applied to solve the problem of learning representations of the multi-view data. Micro-videos are multi-view data, containing visual, acoustic and textual modalities. We thereby introduce the parallel LSTMs to represent each modality of a micro-video as a fixed length of vector, and then we map the vector representations of multiple modalities into a common space with the same length.

3.3.1 Parallel LSTMs. We use $\{\boldsymbol{e}(x^m_{i,1}), ..., \boldsymbol{e}(x^m_{i,N})\}$ to represent the features extracted from sequential units in each modality, where $\boldsymbol{e}(x^m_{i,n})$ denotes the feature for the n-th unit of the i-th micro-video, and $m \in \{v, a, t\}$ denotes the visual modality v, acoustic modality a, or textual modality t.

The features are then fed into parallel LSTMs. At each time step n, LSTM takes the vector $\boldsymbol{e}(x^m_{i,n})$, hidden state vector $\boldsymbol{h}^m_{i,n-1}$, and memory cell vector $\boldsymbol{C}^m_{i,n-1}$ as inputs, and updates $\boldsymbol{h}^m_{i,n}$ and $\boldsymbol{C}^m_{i,n}$ as follows,

$$\begin{cases} \boldsymbol{in}^m_{i,n} = \sigma(\boldsymbol{W}^m_i \boldsymbol{e}(x^m_{i,n}) + \boldsymbol{U}^m_i \boldsymbol{s}^m_{i,n-1} + \boldsymbol{b}^m_i) \\ \boldsymbol{f}^m_{i,n} = \sigma(\boldsymbol{W}^m_f \boldsymbol{e}(x^m_{i,n}) + \boldsymbol{U}^m_f \boldsymbol{s}^m_{i,n-1} + \boldsymbol{b}^m_f) \\ \boldsymbol{o}^m_{i,n} = \sigma(\boldsymbol{W}^m_o \boldsymbol{e}(x^m_{i,n}) + \boldsymbol{U}^m_o \boldsymbol{s}^m_{i,n-1} + \boldsymbol{b}^m_o) \\ \tilde{\boldsymbol{C}}^m_{i,n} = \tanh(\boldsymbol{W}^m_C \boldsymbol{e}(x^m_{i,n}) + \boldsymbol{U}^m_C \boldsymbol{s}^m_{i,n-1} + \boldsymbol{b}^m_C) \\ \boldsymbol{C}^m_{i,n} = \boldsymbol{f}^m_{i,n} \odot \boldsymbol{C}^m_{i,n-1} + \boldsymbol{in}^m_{i,n} \odot \tilde{\boldsymbol{C}}^m_{i,n} \\ \boldsymbol{s}^m_{i,n} = \boldsymbol{o}^m_{i,n} \odot \tanh(\boldsymbol{C}^m_{i,n}) \end{cases}, \tag{8}$$

where $\boldsymbol{in}^m_{i,n}$, $\boldsymbol{f}^m_{i,n}$, and $\boldsymbol{o}^m_{i,n}$ are the input gate, the forget gate and the output gate, respectively; $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic function, $\odot$ is the element-wise multiplication operator, and $\boldsymbol{W}^m_l$, $\boldsymbol{U}^m_l$, and $\boldsymbol{b}^m_l$ for $l \in \{in, f, o, C\}$ are the parameters for the LSTMs. At n=1, $\boldsymbol{s}^m_{i,0}$ and $\boldsymbol{C}^m_{i,0}$ are initialized as zero.

An attention-based pooling is utilized to generate the final vector by a weighted sum of the sequences of vectors $\{\boldsymbol{s}^m_{i,1}, ..., \boldsymbol{s}^m_{i,N}\}$. This pooling method assigns different weights to the vectors of different units, capturing their relative importance. Formally, the process is defined as:

$$\theta(i,m,n,j) = ReLU(\boldsymbol{W}^m_{\text{att}} \boldsymbol{s}^m_{i,n} + \boldsymbol{U}^m_{\text{att}} \boldsymbol{e}(h_j) + \boldsymbol{b}^m_{\text{att}}), \tag{9}$$

$$\alpha(i,m,n,j) = softmax(\theta(i,m,n,j)) = \frac{\exp\theta(i,m,n,j)}{\sum_{n=1}^{N}\exp\theta(i,m,n,j)}, \tag{10}$$

$$\boldsymbol{s}^m_i = \sum_{n=1}^{N} \alpha(i,m,n,j)\boldsymbol{s}^m_{i,n}, \tag{11}$$

where $\boldsymbol{W}^m_{\text{att}}$ and $\boldsymbol{U}^m_{\text{att}}$ are the weight matrices of the attention network, and $\boldsymbol{b}^m_{\text{att}}$ is the bias vector.

3.3.2 Common Space Learning. The parallel LSTMs output three feature vectors with different lengths. Traditional approaches fuse these features with simple concatenation or feature selection. However, we argue that they may not work well in capturing the semantic of features and may hence lead to information redundancy at the learning stage. Therefore, we project the output of LSTMs into a low-dimension common subspace where it can capture the commonality among all the views by three mapping functions $f^v_{map}(\cdot)$, $f^a_{map}(\cdot)$, and $f^t_{map}(\cdot)$, resulting the visual, acoustic, and textual embedding in the common space:

$$\begin{cases} \tilde{\boldsymbol{e}}(x^v_i) = f^v_{map}(\boldsymbol{s}^v_i) \\ \tilde{\boldsymbol{e}}(x^a_i) = f^a_{map}(\boldsymbol{s}^a_i) \\ \tilde{\boldsymbol{e}}(x^t_i) = f^h_{map}(\boldsymbol{s}^t_i) \end{cases}, \tag{12}$$

where $\tilde{\boldsymbol{e}}(x^v_i)$, $\tilde{\boldsymbol{e}}(x^a_i)$, $\tilde{\boldsymbol{e}}(x^t_i) \in \mathbb{R}^{d_F}$ and d_F is the dimension of the embedding in the common space.

3.4 User Embedding

We model users by analyzing users' behaviors and preferences. In particular, we use the visual features of micro-videos (extracted with pretrained CNN) and textual feature of hashtags (extracted with pretrained Word2Vec) posted by users to represent the user embedding. These features are concatenated after an average pooling and fed into a three-layer fully connected neural network, resulting a user representation $\boldsymbol{e}(u_k) \in \mathbb{R}^{d_E}$, where d_E denotes the dimension of the user embedding.

3.5 Interactive Embedding Model

We employ a multi-layer preceptron network to perform an end-to-end learning on both embeddings and interaction functions. Figure 3 illustrates the interactive embedding model. Specifically, we cast the embeddings of micro-video, hashtag, and user into the Bi-Interaction layer and the hidden layers to predict the score.

3.5.1 Bi-Interaction Layer. The Bi-Interaction layer consists of a pooling operation that converts the embedding vectors into one vector:

$$\boldsymbol{p}_0 = \varphi_{\text{pooling}}(\tilde{\boldsymbol{e}}(x^v_i), \tilde{\boldsymbol{e}}(x^a_i), \tilde{\boldsymbol{e}}(x^t_i), \boldsymbol{e}(h_j), \boldsymbol{e}(u_k)), \tag{13}$$

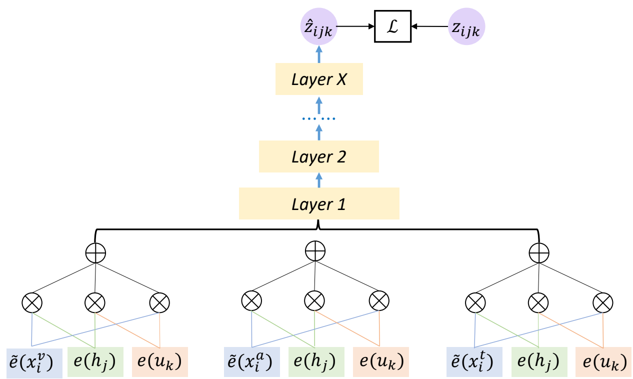


Figure 3: Illustration of the interactive embedding model based on neural network.

$$\varphi_{\text{pooling}} = \begin{bmatrix} \tilde{e}(x_i^v) \odot e(h_j) + \tilde{e}(x_i^v) \odot e(u_k) + e(h_j) \odot e(u_k) \\ \tilde{e}(x_i^a) \odot e(h_j) + \tilde{e}(x_i^a) \odot e(u_k) + e(h_j) \odot e(u_k) \\ \tilde{e}(x_i^t) \odot e(h_j) + \tilde{e}(x_i^t) \odot e(u_k) + e(h_j) \odot e(u_k) \end{bmatrix}, \tag{14}$$

where $\odot$ denotes the element-wise product.

3.5.2 Hidden Layers. The hidden layers consists of fully connected layers, which capture the nonlinear correlations among the micro-videos, hashtags, and users. Formally, they are defined as:

$$\begin{cases} p_1 = ReLU(W_1 p_0 + b_1) \\ p_2 = ReLU(W_2 p_1 + b_2) \\ \ldots\ldots \\ p_X = ReLU(W_X p_{X\text{-}1} + b_X) \end{cases}, \tag{15}$$

where W_X denotes the weight matrix, b_X is the bias vector, and p_X represents the output of the X-th hidden layer.

3.5.3 Prediction Layers. Finally, the output of the last hidden layer p_X is transformed to a prediction score via,

$$\hat{z}_{ijk} = Sigmoid(W_{\text{pre}} p_X), \tag{16}$$

where W_{pre} denotes the weights of the prediction layer. Sigmoid is utilized to regularize the prediction score to the range of [0,1]. An observed interaction is assigned to a target value 1, otherwise 0. To learn the parameters of the neural networks, we optimize the pointwise log loss, as implemented in [17] which forces the prediction score $\hat{z}_{ijk}$ to close to the target z_{ijk} as follows,

$$\begin{aligned} \mathcal{L} &= -\sum_{\tau \in Q^+} log \hat{z}_{ijk} - \sum_{\tau \in Q^-} log(1 - \hat{z}_{ijk}) \\ &= -\sum_{\tau \in Q^+ \cup Q^-} z_{ijk} log \hat{z}_{ijk} + (1 - z_{ijk} log(1 - \hat{z}_{ijk}), \end{aligned} \tag{17}$$

where $\tau = (vi_i, h_j, u_k) \in Q^+$ as a valid interaction if user u_k has added hashtag h_j for his/her posted micro-video vi_i, otherwise $\tau \in Q^-$.

4 EXPERIMENTS

We implemented our method based on PyTorch. We randomly initialized model parameters with Gaussian distribution, and optimized the model with Adam optimizer. The mini-batch size and learning rate were searched in [256; 512; 1024; 2048] and [0:00005; 0:0001; 0:0005; 0:001], respectively. The dimension of the embedding of the visual, acoustic and textual modalities were 500, 300 and 80, respectively. Finally, We set the dimension of the common space as 150. For the correlation matrix, we set α in Equation (5) to be 0.2.

Table 1: Statistics of INSVIDEO.

#(users)	#(videos)	#(hashtags)
6,786	213,847	15,751
#(videos)/user	**#(hashtags)/video**	**average time span**
31.5	13.4	30s

4.1 Dataset

Two public micro-video datasets released by previous studies [1, 45] contain little hashtag information. For example, in [45], each micro-video only has 0.9 hashtag on average. There is no suitable dataset with enough hashtags for our problem. Therefore, we constructed our own dataset INSVIDEO with 213,847 micro-videos and 6,786 users. On average, each micro-video has 13.4 hashtags.

We detailed the dataset construction process as follows. We first crawled micro-videos from the Instagram. In particular, we manually chose hashtags from hashtag dictionary website[2] as our seed hashtags. The hashtags are organized into a four-layer hierarchical structure, with 16, 1,333, and 4,092 leaf nodes in the second-layer, third-layer and fourth-layer, respectively. We then searched the hashtags on Instagram and collected at most the top nine posts for each hashtag. and regarded their users who post these posts as active users. For each active user, we crawled his/her at most 50 published micro-videos and video descriptions. In this way, we harvested 334,826 micro-videos from 9,170 active users.

We further conducted data cleaning on micro-videos, hashtags, and users. For micro-videos, we removed the videos with no hashtags or missing modalities (visual, acoustic and text). For hashtags, we conducted spell checking and word lemmatization, and then removed the hashtags occurring less than 50 times. For users, we removed the users with less than 10 micro-videos. After the data cleaning, we obtained a dataset of 213,847 micro-videos and 15,751 hashtags from 6,786 users and each has 13.4 hashtags on average. Besides, the average time span of the micro-videos is 30s. The statistics of the INSVIDEO are summarized in Table 1.

We also performed an analysis on the frequency of the collected hashtags. As shown in Figure 4, the hashtag frequency distribution is heavily skewed towards a few frequent hashtags and a long-tail of rare hashtags. For example, the most frequent hashtag, #love, appears over 42,000 times in the dataset. The least frequent hashtag only appears 50 times (*e.g.,* light drip). The original dataset was further divided into three separate datasets based on the micro-video instances, with 80%, 10% and 10% randomly for training, validation and testing, respectively. The hashtags are kept as the same for these three sets. Moreover, we randomly sampled 6 negative hashtags to pair with each positive instance.

[2] https://tagsforlikes.com/

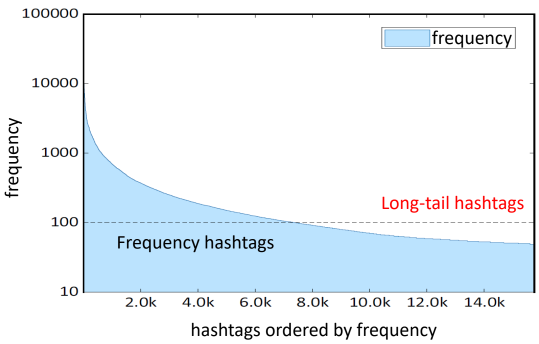


Figure 4: Hashtag frequency distribution in our collected INSVIDEO dataset.

4.2 Feature Extraction

We converted micro-videos to frame sequences with FFmpeg[3] and selected 40 frames for each micro-video with uniform sampling. For each frame, we extracted a 2,048 dimensional feature vector with a pretrained ResNet [16] on ImageNet. We adopted Librosa[4] to extract a 128-dimensional feature vector for each 0.2s audio clip. We uniformly sampled 60 acoustic feature vectors for each micro-video. With the video description, we first removed the non-English words and stop words, and performed word lemmentation. We then employed Word2Vec [30] to generate vector representation for words. It is noted that we selected at most 6 words for each micro-video.

4.3 Experimental Setting

4.3.1 Evaluation Metrics. Given a micro-video in the testing set, our method outputs prediction scores for all hashtags to rank them accordingly. We predicted top-K hashtags for each of the test micro-videos and compared it with the ground truth. To evaluate the performance, we employed the widely used metrics: Recall@K and NDCG@K. Recall@K measures whether the item of the ground truth is in the predicted top-K list, while NDCG@K accounts for the position of hit by assigning a larger score to the higher position. Following the commonly used setting in recommendation systems, we used K=5 and K=10 in our experiment.

4.3.2 Long-tail Recommendation. To evaluate the effect of the recommendation algorithm on the long-tail hashtags, we constructed a new testing set which only contains long-tail hashtags. By investigating the dataset, we treated the hashtags which appear less than 100 times as long-tail hashtags, and treated the others as frequent hashtags. Specifically, we modified the prior testing set by removing the frequent hashtags in the ground truth. Similarly, we employed Recall@K and NDCG@K as our evaluation metrics.

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

ConTagNet [33]. This is a CNN-based method which integrates context with image content for multi-lable tag prediction. The method considers both the visual information and the context in which the photo has been captured, such as time and location. We adopted the released implementation[5], with only the visual information considering that there is no context in our dataset.

Co-Attention [46]. This is the state-of-the-art hashtag recommendation method in Twitter. It introduces a co-attention network, incorporating both the textual and visual information. We employed the implementation released by the authors[6].

User-specific Hashtag Modeling [37]. This is a three-way tensor model which is responsible for modeling the interactions among image features, hashtag embeddings, and user embeddings. We implemented it by replacing the image features with the visual, acoustic, and textual features of the micro-videos.

V2HT$^{\text{w/o UP}}$, V2HT$^{\text{w/o P}}$, V2HT$^{\text{w/o U}}$. These are variants of V2HT method by removing the propagation module and user module (V2HT$^{\text{w/o UP}}$), propagation module (V2HT$^{\text{w/o P}}$), and user module (V2HT$^{\text{w/o U}}$) to demonstrate the effect of the propagation mechanism and the video-hashtag-user interaction learning.

Table 2: Performance comparison between baselines and our proposed method with its variants.

Methods	K=5		K=10	
	Recall	NDCG	Recall	NDCG
ContagNet	0.3996	0.2919	0.4961	0.3232
Co-attention	0.4276	0.3109	0.5307	0.3444
User-Specific	0.5063	0.3809	0.6304	0.4211
V2HT$^{\text{w/o UP}}$	0.4726	0.3516	0.5748	0.3837
V2HT$^{\text{w/o P}}$	0.5560	0.4290	0.6659	0.4647
V2HT$^{\text{w/o U}}$	0.4821	0.3637	0.5841	0.3963
V2HT	**0.6166**	**0.5236**	**0.6948**	**0.5489**

4.4 Results and Discussion

4.4.1 Overall Performance Comparison. Experimental results of the comparison between baselines and our proposed method with its variants are summarized in Table 2. We have the following observations: First, our V2HT model achieves the best performance on both Recall and NDCG, and significantly outperforms other state-of-the-art methods. Second, compared to user-agnostic hashtag model Co-attention and ContagNet, User-specific achieves obvious improvement. The trend is similar on our proposed method that V2HT and V2HT$^{\text{w/o P}}$ have better performance compared to V2HT$^{\text{w/o U}}$ and V2HT$^{\text{w/o UP}}$, respectively. The reason is that the user embedding module encodes the user's preferences, which is essential to be considered. Third, after adding the hashtag propagation mechanism, the performance improves 0.95% in Recall@5 and 1.21% in NDCG@5 (V2HT$^{\text{w/o U}}$ vs. V2HT$^{\text{w/o UP}}$), and 6.06% in Recall@5 and 9.46% in NDCG@5 (V2HT$^{\text{w/o P}}$ vs. V2HT). It verifies the effectiveness of our proposed propagation mechanism. It is interesting to note that adding user information alone is more useful than adding propagation mechanism alone (V2HT$^{\text{w/o P}}$ vs. V2HT$^{\text{w/o U}}$), however, the usage of propagation mechanism is more prominent with the presence of user module.

[3]https://www.ffmpeg.org
[4]https://github.com/librosa
[5]https://github.com/vyzuer/contagnet
[6]http://jkx.fudan.edu.cn/~qzhang/paper/code/IJCAI2017.zip

Table 3: Experimental results on recommending long tail hashtags.

Methods	K=5		K=10	
	Recall	NDCG	Recall	NDCG
ContagNet	0.0367	0.0114	0.1039	0.0329
Co-attention	0.0374	0.0175	0.1109	0.0409
User-Specific	0.0528	0.0288	0.1414	0.0573
V2HT$^{\text{w/o UP}}$	0.0918	0.0491	0.1948	0.0821
V2HT$^{\text{w/o P}}$	0.0571	0.0313	0.1535	0.0621
V2HT$^{\text{w/o U}}$	**0.1217**	**0.0719**	**0.2454**	**0.1118**
V2HT	0.0590	0.0358	0.1641	0.0697

4.4.2 Performance Comparison during Training. We further analyzed the learning trend of our proposed methods and reported it in Figure 5. During the convergence process, we observed that after adding the propagation mechanism, the V2HT and V2HT$^{\text{w/o U}}$ consistently outperforms V2HT$^{w/oP}$ and V2HT$^{\text{w/o UP}}$, respectively. It demonstrates that the model with the propagation mechanism can speed up the convergence and achieve better results. We further noticed that the proposed methods with user module (V2HT and V2HT$^{\text{w/o P}}$) are inferior to that without user module (V2HT$^{\text{w/o U}}$ and V2HT$^{\text{w/o UP}}$) at the early stage. However, the methods with user module achieve better results at the late stage. Though user information (showing users' hashtag usage patterns) will increase the complexity of the model, we still believe it is an important factor in recommending hashtags for users.

4.4.3 Evaluation on Long-tail Hashtag Recommendation. We used the protocol in Section 4.3.2 to evaluate the long-tail recommendation and showed the results in Table 3. From the results we can see that, in general, all the performance is inferior to that on the regular dataset, and our proposed V2HT and its variants outperform all the other state-of-the-art methods on this long-tail hashtag sub-dataset. Among all the V2HTs, V2HT$^{\text{w/o U}}$ achieves the best performance. This is not surprising since we have also seen a bigger improvement when adding the user embedding module on regular dataset as shown in Table 2. The influence of user embedding and hashtag propagation may contract to certain extent on the selection of long-tail hashtag. However, we still believe that the proposed hashtag propagation mechanism is useful given the significant improvement (32.6%, 46.4%, 26.0%, and 36.2% for Rec@5, NDCG@5, Rec@10, and NDCG@10, respectively) from V2HT$^{\text{w/o UP}}$ to V2HT$^{\text{w/o U}}$.

4.4.4 Evaluation on Modality Combination. To demonstrate the usage of multi-modal data for hashtag recommendation, we performed the study on V2HT by replacing micro-video embedding with various modality combinations. We have the following observations: 1) In terms of the single modality comparison, *Visual* significantly outperforms *Acoustic* and *Textual*. This is mainly because the visual modality provides primary information of micro-videos and thus promotes the hashtag performance. 2) In terms of the modality combinations, the more modalities are considered in the model, the better performance can be achieved. It verifies the assumption that the different modalities are complementary to each other. And 3) *Visual+Acoustic+Text* achieves the best performance. This validates the effectiveness in aggregating multiple modalities of our V2HT framework. In addition, the performance trend on multimodal data integration is the same on the comparison between our proposed methods with other baselines (as shown in Table 2) that V2HT$^{\text{w/o UP}}$ (with visual, textual, and acoustic information) outperforms the Co-attention (with visual and textual information), and Co-attention outperforms the visual modality only method ContagNet.

Table 4: Overview performance comparison of various methods.

Methods	K=5		K=10	
	Recall	NDCG	Recall	NDCG
Textual	0.4225	0.3440	0.5028	0.3699
Acoustic	0.4919	0.4206	0.5635	0.4437
Visual	0.5389	0.4724	0.6035	0.4933
Acoustic+Textual	0.5312	0.4698	0.6017	0.4788
Visual+Textual	0.5892	0.5005	0.6490	0.5198
Visual+Acoustic	0.6087	0.5172	0.6711	0.5374
Visual+Acoustic+Text	**0.6166**	**0.5236**	**0.6948**	**0.5489**

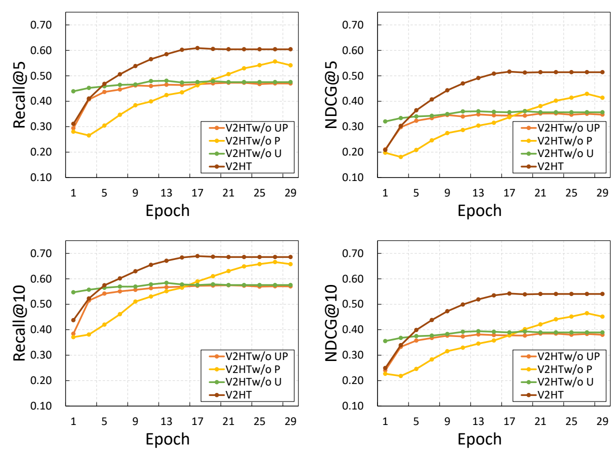


Figure 5: Experimental results of the comparison between our proposed methods and its variants during training.

4.5 Case study

In order to achieve a deeper understanding of what hashtags are recommended by our proposed model, we presented a qualitative analysis of three case studies. We selected three representative types of micro-video (*i.e.,* singing, sports and dance) in our dataset, and presented their ground truth hashtags and the hashtags predicted by our proposed methods in Figure 6.

From the first example of singing scenario, we can see that the methods with user embedding module predicted more personalized hashtags (*e.g.,* #bangerz tour and #rip hannah montana), which might be brought by the knowledge from user's previous posted videos or hashtags. We have also noticed a positive effect on the propagation mechanism on example (b) that two long-tail hashtags (*e.g.,* #viral dance appears 77 times, and #kid dancer appears 68

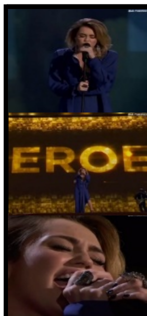
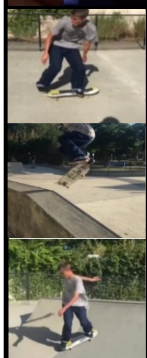
(a)		**Ground truth Hashtags**	bangerz tour, love, bangerz, miler, malibu, f bangerz tour, fashion, hannah montana, gain post, cant stop, rip hannah montana, cant tame, spam, nothing break like heart, young, miley cyrus, like, wreck ball, edit
		V2HT$^{w/o\ UP}$	music, ***love***, good, ***like***, video, dance, hiphop, daily, nicki minaj, family
		V2HT$^{w/o\ U}$	***love***, ***like***, music, reputation, style, taylor swift, ariana grande, beautiful, good, singer
		V2HT$^{w/o\ P}$	good, ***love***, ***f bangerz tour***, ***miler***, dance, music, ***rip hannah montana***, selena gomez, ***like***, cant stop
		V2HT	***f bangerz tour***, ***rip hannah montana***, ***cant tame***, taylor swift, ***bangerz tour***, ***miley cyrus***, selena gomez, ariana grande, beyonce, ***wreck ball***
(b)		**Ground truth Hashtags**	skate, skate clip, apl, trendy squad, trendy, skate crunch, skate die, trend skate, skate shop, love skateboard, skate damn day, skate life, skater boy, skater girl, skateboard fun, metro, skateboard, clip, skatepark, video day, skate fam, skate spot, ber ric, gucci, adidas
		V2HT$^{w/o\ UP}$	skateboard crime, ber ric, ***skateboard fun***, ***skateboard***, ***skate life***, skateboarder, ***skate damn day***, skater, sker, skate
		V2HT$^{w/o\ U}$	***skateboard***, ***skate life***, ***metro***, ***skate***, ***skateboard fun***, ***skate damn day***, ***skate crunch***, ***skatepark***, skater, ***skate spot***
		V2HT$^{w/o\ P}$	***trend skate***, ***skateboard***, ***skateboard fun***, training, fit, ***skate life***, ***trendy squad***, ***apl***, muscle, ***skate damn day***
		V2HT	***trend skate***, ***trendy squad***, ***apl***, ***skate clip***, ***skate shop***, ***skate fam***, ***skate spot***, ***skater girl***, skate clip daily, ***skate die***
(c)		**Ground truth Hashtags**	trend, hiphop, viral dance, dance renaissance, kid dancer, dance, explore page, good, viral video, viral, jazz, dancer, hiphop dance
		V2HT$^{w/o\ UP}$	dance class, ***dance***, choreographer, ***dancer***, love dance, music, fitness, choreography, hiphop dancer, dance studio
		V2HT$^{w/o\ U}$	***dancer***, ***hiphop***, ***dance***, choreography, ***viral dance***, ***hiphop dance***, ***kid dancer***, love, music, dance class
		V2HT$^{w/o\ P}$	***dance renaissance***, dance life, ***dance***, ***dancer***, ***hiphop***, dance class, music, fitness, ***good***, girl
		V2HT	***dance renaissance***, ***kid dancer***, ***viral dance***, ***hiphop dance***, ***dancer***, dance challenge, dance video, ***good***, ***hiphop***, fitness

Figure 6: Case study for three representative micro-video scenarios. For each example, the selected three snapshots, ground truth hashtags posted by users, and predicted hashtags by V2HT$^{w/o\ UP}$, V2HT$^{w/o\ U}$, V2HT$^{w/o\ P}$, and V2HT are presented.

times) are predicted by V2HT and V2HT$^{w/oU}$ with the hashtag propagation mechanism included.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a multi-view representation interactive embedding model with graph-based information propagation for micro-video hashtag recommendation. It considers the multi-view learning, the hashtag correlations, and the video-user-hashtag interaction simultaneously. In particular, we construct a graph to guide the information propagation process among hashtags. By leveraging the predefined structure to regularize the relatedness among hashtags, the hashtag recommendation performance has a significant improvement on both frequent and long-tail hashtags. The experiment results demonstrate our proposed method achieves the state-of-the-art performance for the hashtag recommendation.

In the future, we plan to extend our work in the following two directions. First, we plan to introduce attention mechanism into interactive embedding model to focus on the important cues among multimodal features [9–11] of micro-videos, hashtags and users. Second, we expect to reduce redundant hashtags for micro-videos. Third, we would like to work on the explainability of hashtag recommendation [3–5, 23].

6 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.:61702300, No.:61702302, No.: 61802231, and No. U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.: ZR2019JQ23, No.:ZR2019QF001; the Future Talents Research Funds of Shandong University, No.: 2018WLJH 63.

REFERENCES

[1] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. In *Proceedings of ACM Conference on Multimedia Conference*. 898–907.

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition with Graph Convolutional Networks. *arXiv preprint arXiv:1904.03582* (2019).

[3] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose Catherine Kanjirathinkal, and Mohan S. Kankanhalli. [n. d.]. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Transactions on Information Systems* ([n. d.]).

[4] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S. Kankanhalli. 2018. A^3NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3748–3754.

[5] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan S. Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In *The World Wide Web Conference*. 639–648.

[6] Emily Denton, Jason Weston, Manohar Paluri, Lubomir D. Bourdev, and Rob Fergus. 2015. User Conditional Hashtag Prediction for Images. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1731–1740.

[7] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning Topical Translation Model for Microblog Hashtag Suggestion. In *Proceedings of International Joint Conference on Artificial Intelligence*. 2078–2084.

[8] Miles Efron. 2010. Hashtag retrieval in a microblogging environment. In *Proceeding of ACM SIGIR Conference on Research and Development in Information Retrieval*. 787–788.

[9] Tian Gan, Junnan Li, Yongkang Wong, and Mohan S. Kankanhalli. 2019. A Multi-sensor Framework for Personal Presentation Analytics. *Transaction on Multimedia Computing, Communications, and Applications* 15, 2 (2019), 30:1–30:21.

[10] Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, and Mohan S. Kankanhalli. 2015. Multi-sensor Self-Quantification of Presentations. In *ACMMM*. 601–610.

[11] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S. Kankanhalli. 2013. Temporal encoded F-formation system for social interaction detection. In *Proceedings of ACM International Conference on Multimedia*. 937–946.

[12] Yuqi Gao, Jitao Sang, Tongwei Ren, and Changsheng Xu. 2017. Hashtag-centric Immersive Search on Social Media. In *Proceedings of ACM on Multimedia Conference*. 1924–1932.

[13] Yuyun Gong and Qi Zhang. 2016. Hashtag Recommendation Using Attention-Based Convolutional Neural Network. In *Proceedings of International Joint Conference on Artificial Intelligence*. 2782–2788.

[14] Yeyun Gong, Qi Zhang, and Xuanjing Huang. 2015. Hashtag Recommendation Using Dirichlet Process Mixture Models Incorporating Types of Hashtags. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. 401–410.

[15] Elaheh Malekzadeh Hamedani and Marjan Kaedi. 2019. Recommending the long tail items through personalized diversification. *Knowledge Based Systems* (2019), 348–357.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of International Conference on World Wide Web*. 173–182.

[18] Yu-Chieh Ho, Yi-Ting Chiang, and Jane Yung-jen Hsu. 2014. Who likes it more?: mining worth-recommending items from long tails by modeling relative preference. In *ACM International Conference on Web Search and Data Mining*. 253–262.

[19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

[20] Svetlana Kordumova, Jan C. van Gemert, and Cees G. M. Snoek. 2016. Exploring the Long Tail of Social Media Tags. In *MultiMedia Modeling*. 51–62.

[21] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. 2012. On Recommending Hashtags in Twitter Networks. In *Social Informatics*. 337–350.

[22] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-Label Zero-Shot Learning With Structured Knowledge Graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1576–1585.

[23] Chenliang Li, Cong Quan, Li Peng, Yunwei Qi, Yuming Deng, and Libing Wu. 2019. A Capsule Network for Recommendation and Explaining What You Like and Dislike. *Computing Research Repository* (2019).

[24] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. 2016. Conditional Graphical Lasso for Multi-label Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2977–2986.

[25] Xin Li, Feipeng Zhao, and Yuhong Guo. 2014. Multi-label Image Classification with A Probabilistic Label Enhancement Model. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*. 430–439.

[26] Yang Li, Ting Liu, Jing Jiang, and Liang Zhang. 2016. Hashtag Recommendation with Topical Attention-Based LSTM. In *International Conference on Computational Linguistics*. 3019–3029.

[27] Jie Liu, Zhicheng He, and Yalou Huang. 2018. Hashtag2Vec: Learning Hashtag Representation with Relational Hierarchical Embedding Model. In *Proceedings of International Joint Conference on Artificial Intelligence*. 3456–3462.

[28] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards Micro-video Understanding by Joint Sequential-Sparse Modeling. In *Proceedings of the ACM International Conference on Multimedia*. 970–978.

[29] Dhruv Mahajan, Vishwajit Kolathur, Chetan Bansal, Suresh Parthasarathy, Sundararajan Sellamanickam, S. Sathiya Keerthi, and Johannes Gehrke. 2016. Hashtag Recommendation for Enterprise Applications. In *Proceedings of ACM Conference on Information and Knowledge Management*. 893–902.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Annual Conference on Neural Information Processing Systems*. 3111–3119.

[31] Minseok Park, Hanxiang Li, and Junmo Kim. 2016. HARRISON: A Benchmark on HAshtag Recommendation for Real-world Images in Social Networks. *Computing Research Repository* abs/1605.05054 (2016).

[32] Yoon-Joo Park. 2013. The Adaptive Clustering Method for the Long Tail Problem of Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering* 25, 8 (2013), 1904–1915.

[33] Yogesh Singh Rawat and Mohan S. Kankanhalli. 2016. ConTagNet: Exploiting User Context for Image Tag Recommendation. In *Proceedings of ACM Conference on Multimedia Conference*. 1102–1106.

[34] Ruslan Salakhutdinov, Antonio Torralba, and Joshua B. Tenenbaum. 2011. Learning to share visual appearance for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1481–1488.

[35] Bidisha Samanta, Abir De, Abhijnan Chakraborty, and Niloy Ganguly. 2017. LMPP: A Large Margin Point Process Combining Reinforcement and Competition for Modeling Hashtag Popularity. In *Proceedings of International Joint Conference on Artificial Intelligence*. 2679–2685.

[36] Lei Shi. 2013. Trading-off among accuracy, similarity, diversity, and long-tail: a graph-based recommendation approach. In *ACM Conference on Recommender Systems*. 57–64.

[37] Andreas Veit, Maximilian Nickel, Serge J. Belongie, and Laurens van der Maaten. 2018. Separating Self-Expression and Visual Content in Hashtag Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5919–5927.

[38] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of ACM Conference on Information and Knowledge Management*. 1031–1040.

[39] Yilin Wang, Suhang Wang, Jiliang Tang, Guo-Jun Qi, Huan Liu, and Baoxin Li. 2017. CLARE: A Joint Approach to Label Classification and Tag Recommendation. In *Proceedings of AAAI Conference on Artificial Intelligence*. 210–216.

[40] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural Multimodal Cooperative Learning Towards Micro-video Understanding. *IEEE Transactions on Image Processing* (2019).

[41] Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. 2009. Learning to tag. In *Proceedings of International Conference on World Wide Web*. 361–370.

[42] Zhibiao Wu and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of Annual Meeting on Association for Computational Linguistics*. 133–138.

[43] Chen Xing, Yuan Wang, Jie Liu, Yalou Huang, and Wei-Ying Ma. 2016. Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter. In *Proceedings of AAAI Conference on Artificial Intelligence*. 2666–2672.

[44] Toshihiko Yamasaki, Jiani Hu, Shumpei Sano, and Kiyoharu Aizawa. 2017. FolkPopularityRank: Tag Recommendation for Enhancing Social Popularity using Text Tags in Content Sharing Services. In *Proceedings of International Joint Conference on Artificial Intelligence*. 3231–3237.

[45] Jianglong Zhang, Liqiang Nie, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2016. Shorter-is-Better: Venue Category Estimation from Micro-Video. In *Proceedings of ACM Conference on Multimedia Conference*. 1415–1424.

[46] Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. 2017. Hashtag Recommendation for Multimodal Microblog Using Co-Attention Network. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3420–3426.

[47] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. 2014. Capturing Long-Tail Distributions of Object Subcategories. In *IEEE Conference on Computer Vision and Pattern Recognition*. 915–922.

[48] Xiaoyu Zhu, Tian Gan, Xuemeng Song, and Zhumin Chen. 2017. Sentiment Analysis for Social Sensor. In *Advances in Multimedia Information Processing, Pacific-Rim Conference on Multimedia*. 893–902.