Personalized Hashtag Recommendation for Micro-videos

Yinwei Wei
Shandong University
weiyinwei@hotmail.com

Zhiyong Cheng
Shandong Computer Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academic of Sciences)
jason.zy.cheng@gmail.com

Xuzheng Yu
Shandong University
XuzhengYuuu@gmail.com

Zhou Zhao
Zhejiang University
zhaozhou@zju.edu.cn

Lei Zhu
Shandong Normal University
leizhu0608@gmail.com

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

ABSTRACT

Personalized hashtag recommendation methods aim to suggest users hashtags to annotate, categorize, and describe their posts. The hashtags, that a user provides to a post (e.g., a micro-video), are the ones which *in her mind* can well describe *the post content where he/she is interested in*. It means that we should consider both *users' preferences on the post contents* and *their personal understanding on the hashtags*. Most existing methods rely on modeling either the interactions between hashtags and posts or the interactions between users and hashtags for hashtag recommendation. These methods have not well explored the complicated interactions among users, hashtags, and micro-videos. In this work, towards the personalized micro-video hashtag recommendation, we propose a Graph Convolution Network based Personalized Hashtag Recommendation (GCN-PHR) model, which leverages recently advanced GCN techniques to model the complicate interactions among <users, hashtags, micro-videos> and learn their representations. In our model, the users, hashtags, and micro-videos are three types of nodes in a graph and they are linked based on their direct associations. In particular, the message-passing strategy is used to learn the representation of a node (e.g., user) by aggregating the message passed from the directly linked other types of nodes (e.g., hashtag and micro-video). Because *a user is often only interested in certain parts of a micro-video* and *a hashtag is typically used to describe the part (of a micro-video) that the user is interested in*, we leverage the attention mechanism to filter the message passed from micro-videos to users and hashtags, which can significantly improve the representation capability. Extensive experiments have been conducted on two real-world micro-video datasets and demonstrate that our model outperforms the state-of-the-art approaches by a large margin.

*Liqiang Nie and Zhiyong Cheng are the corresponding authors.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/

CCS CONCEPTS

• **Engaging users with multimedia → Multimedia Search and Recommendation**.

KEYWORDS

Graph Neural Network, Personalized Hashtag Recommendation, Micro-video Understanding

ACM Reference Format:
Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized Hashtag Recommendation for Micro-videos. In *Processdings of the 27th ACM Int'l Conf. on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/

1 INTRODUCTION

In social networks, such as Twitter[1] and Instagram[2], hashtags are frequently used to annotate, categorize, and describe the posts according to users' preferences. They may consist of any arbitrary combination of characters led by a hash symbol '#' (e.g. #Puppy and #thegoodlife). Hashtags are created by users, and they hence can be treated as the self-expression of users, conveying the users' preferences on posts and their usage styles of hashtags. With these hashtags, users can easily search and manage their historical posts and track others' posts. Moreover, studies have shown that hashtags can provide valuable information about several tasks, such as sentiment analysis [39] and video understanding [35]. However, due to the inconvenient typing on smartphones, only sparse users are willing to provide hashtags to their posts (e.g. micro-videos). According to the statistics[3], by September 2018, more than 33 million micro-videos without any hashtag are uploaded per day on Instagram. Therefore, personalized hashtag recommendation has attracted considerable attention from industrial and academic communities.

In recent years, several methods have been proposed to automatically suggest appropriate hashtags to users rely on their posts' content. Pioneer efforts [10, 11, 32, 33] view personalized hashtag recommendation as a multi-class classification or information retrieval problem to predict the hashtag. For instance, Tran *et al.* [33] presented a hashtag recommendation method which leverages the

[1] https://twitter.com/.
[2] https://www.instagram.com/.
[3] https://www.omnicoreagency.com/instagram-statistics/.

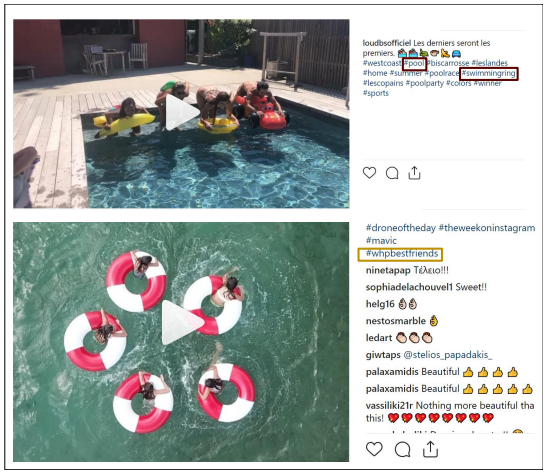

Figure 1: Exemplar demonstration of user preference. Two users publish similar micro-videos with different hashtags.

historical tweets, used hashtags, and social interaction to profile the users, then searched the similar users' tweets for extracting the hashtags. Denton *et al.* [11] used metadata (e.g., age, gender, etc.) to characterize users, and integrated users' representations with features of their posts to infer the categories. Nevertheless, these methods ignore the interactions between users and hashtags. For example, the hashtag '#rock' may be annotated to totally different images by music fans and mountaineers. Users indeed always have their own preferences on hashtag usages. Being aware of this problem, Alvari [1] applied the matrix factorization based collaborative filtering (CF) method to model the interactions between users and hashtags for personalized hashtag recommendation. Furthermore, considering the inherent subjective of the hashtag, Veit *et al.* [34] trained a user-specific hashtag model on image-user-hashtag triplets, by taking the hashtag usage patterns of users into account. However, this method has not well exploited the post content, which contains rich information about user preference and hashtag semantics.

Although the CF-based methods linearly model interactions among users, hashtags, and posts, they merely capture the user-specific hashtag usage patterns rather than the representation of the user preference and hashtag semantic, while the latter is the core of personalized hashtag recommendation. The hashtag semantic should be consistent with the corresponding part of post content, which in turn is the interested part of the user. For example, as shown in Figure 1, different users may annotate different hashtags for similar micro-video according to their preferences. Considering this fact, the user preference and hashtag semantic representation have complicated interaction with posts' features. Therefore, how to exploit such complicated interactions to represent the user preference and hashtag semantic is nontrivial. Especially, the abundant and multimodal information of micro-videos makes hashtag personalized recommendation even more challenging [25, 26, 29]. To deal with the aforementioned challenge, we propose a novel graph convolutional network (GCN) based method to model such complicated interactions for micro-video hashtag personalized recommendation, as illustrated in Figure 2. In our model, the users, hashtags, and micro-videos are three types of nodes in a graph and they are linked based on their direct associations. In other words, user nodes are connected to their historical micro-video nodes and used hashtag nodes, and hashtag nodes are connected to their accompanied micro-video nodes and corresponding user nodes. Based on this graph, our model represents the user nodes and hashtag nodes using the graph convolutional techniques.

In particular, based on the message-passing idea [3], the representation of a node can be learned based on the message passed from its neighbor nodes. In our model, the representation of a user node is learned by aggregating the message from her neighbor hashtag and micro-video nodes; and the representation of a hashtag is learned based on the message from its neighbor user and micro-video nodes. With the observations of that, a user is often only interested in certain parts of a micro-video and a hashtag is typically used to describe a particular part of a micro-video, we deem that the message passed from a micro-video to its neighbor user or hashtag is redundant. To solve this problem, we employ the attention mechanism to filter the information related to the user and hashtag. Specifically, our model uses the hashtag representation to filter the micro-video information in the corresponding user node representation, since hashtags are used by users to express their interests in the micro-video. Analogous, the representation of user preferences are used to filter the micro-video information to model the hashtag semantics, because the user preferences can be used to identify which parts in the micro-video are tagged by the hashtags. Based on this design, our model can achieve better user and hashtag representation learning. Thereafter, our model future learns the user-specific micro-video features and user-specific hashtag semantics with obtained representations of user preference and hashtag semantic, which are then used for personalized hashtag recommendation for micro-videos. To verify the effectiveness of our model, we perform extensive experiments on two micro-video benchmark datasets. The experimental results show that our proposed model outperforms several state-of-the-art approaches.

The main contributions of this work are threefold:

- To the best of our knowledge, this is the first work which attempts to design a personalized hashtag recommendation method for micro-videos. Our model can comprehensively model the interactions between users, hashtags, and micro-videos for hashtag recommendation.
- We design a novel GCN based hashtag recommendation method. In particular, we introduce the attention mechanism to filter the redundant message passed from micro-videos to users and hashtags in the graph, which can significantly enhance the learning of user and hashtag representations.
- We conduct extensive experiments on two real-world micro-video datasets demonstrate the superiority of our method over several state-of-the-art methods. In addition, we released our codes, datasets, and parameters to inspire other researchers[4].

2 RELATED WORK

In this section, we mainly review the studies that are most related to our work, including hashtag recommendation and graph convolutional network.

[4]https://github.com/weiyinwei/GCN_PHR.

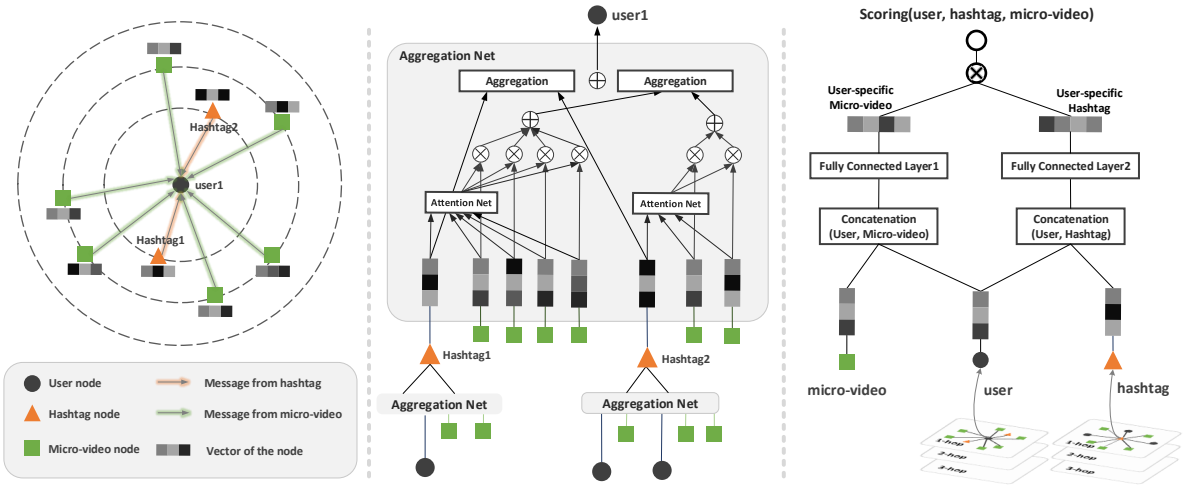

Figure 2: Schematic illustration of our proposed graph-based convolutional network. In this graph, we treat users, hashtags and micro-videos as nodes. We use two edges to link them if a user annotates a micro-video use a hashtag.

2.1 Hashtag Recommendation

Due to the widespread use of hashtags on various social platforms, hashtag recommendation has attracted increasing attention in recent years and many methods [40, 41] have been proposed. Those methods can be broadly grouped into two categories: 1) the ones based on the interactions between hashtags and posts; and 2) the ones based on the interactions between users and posts.

Methods in the first category attempt to learn the representation of hashtags based on their corresponding posts, and then recommend hashtags according to the post content. For example, Zhang *et al.* [45] assumed that hashtag and textual content are two different descriptions of the same thing. Based on that, they proposed a topical translation model, which extracts topics from contents and translates them to the hashtags, to recommend hashtags for microblogs. Dey *et al.* [12] regarded the hashtag recommendation as the word prediction task. Specifically, a hashtag is treated as a word in tweets and then the word-embedding techniques are applied to learning the representation of hashtags and predicting hashtags in tweets. Nevertheless, abovementioned hashtag recommendation methods ignore the user-related factors while recommending hashtags, such as user preference and language habit. In fact, some users may use different hashtags to describe the same content (i.e., synonyms) and some may use the same hashtag to describe different contents (i.e., polysemy). Taking users into account, Veit *et al.* [34] applied the three-way tensor product to learning the user-specific representations for hashtags. The learned representations are then used to measure the similarity scores between hashtags and posts. However, this method merely captures the user preferences on hashtags while ignores user interest in the post content. Different from these methods, our model learns both the user-specific hashtag representation and user-specific post representations, which are then used to compute the user-specific similarity score of each post and hashtag pair.

Methods in the second group treat the hashtag recommendation as a classification or prediction problem by learning the interactions between the user preference and post content. Denton *et al.* [11] took the user demographic information into consideration (e.g., age, gender, etc.) to model user preferences, which are combined with image features and then mapped into the same embedding space with hashtags. In this space, the hashtags are matched with the user-image combination representations for the recommendation. Rawat *et al.* [32] fed the integration of user representation and image features into a deep neural network to predict hashtags for images. In this model, they considered the contextual preferences of users on images (e.g. time and geo-location) in user representation learning. Recently, Park *et al.* [10] modeled user preferences based on their most frequently used hashtags in historical posts and devised a Context Sequence Memory Network for hashtag prediction based on user preferences and image features. All the above methods model user preferences on the global level of a post, however, it is common that a user is interested in a particular part of a post and provides a hashtag according to the interested part, especially for the posts with rich content (such as micro-videos). In this work, we will model the user preference based on the content interests them in micro-videos and learn user-specific micro-video representations according to content that interests them.

2.2 Graph Convolutional Networks

Graph convolutional networks have been widely used in various applications, such as computer vision [14, 20, 23, 27, 44], disease or drug prediction [22, 31, 46], and Chemistry [13, 15, 24, 43]. Due to their powerful capability in representation learning, GCNs have also been exploited to model the interactions between users and items [21, 28, 30, 37, 38] for recommendation. For example, a general inductive GCN framework GraphsSAGE [16], which learns node representation based on both the topological structure of graph and node feature information, has been verified in citation recommendation and video recommendation. However, GCN-based recommendation models often need to store the whole graph in GPU memory, which limits their applications in large-scale recommendation tasks with billions of items and millions of users. To tackle this issue, Ying *et al.* [42] proposed an efficient GCN algorithm PinSage which combines efficient random walks and

graph convolutions to learn node representations. This method has been demonstrated to be scalable for large-scale recommendation tasks. From a different perspective, Berg *et al.* [3] considered recommendation as link prediction on graphs and proposed a novel graph auto-encoder framework named Graph Convolutional Matrix Completion (GCMC). Based on a proposed message-passing model, this method uses a graph convolution layer in the encoder to learn the representations of users and items and then leverages the representations to reconstruct the rating links through a bilinear decoder.

Different from the previous works which typically consider two types of nodes, we need to model the interactions among three kinds of nodes (i.e., user, hashtag, micro-video) in this work. As in GCMC, we also adopt the message-passing model to learn the interaction among the different types of nodes. The difference is that we propose to use the attention mechanism to filter the message from micro-videos to users and hashtags. To the best of our knowledge, this is the first work to use GCN-model for personalized hashtag recommendation and also the first attempt to apply GCN techniques for graphs with three types of nodes.

3 OUR PROPOSED METHOD

3.1 Problem Setting and Model Overview

3.1.1 Problem setting. Before describing our method, we would like to introduce the problem setting first. Given a dataset with a micro-video set $\mathcal{V}$, a hashtag set $\mathcal{H}$, and a user set $\mathcal{U}$, in which a micro-video $v_k \in \mathcal{V}$ is uploaded by a user $u_i \in \mathcal{U}$ and some hashtags $h_j \in \mathcal{H}$ are provided by u_i to v_k. Based on this dataset, the goal is to learn a personalized hashtag recommendation model, which could automatically recommend hashtags from $\mathcal{H}$ to a new micro-video v uploaded by a user u. Because we would like that the recommended hashtags will be adopted by the user, the recommended hashtags should not only match the micro-video contents but also fit the personal preferences of users. To achieve this goal, we apply the graph convolutional networks to modeling the complex interactions among three types of entities: users, hashtags, and micro-videos.

Let $\mathcal{G} = (\mathcal{W}, \mathcal{E})$ be an undirected graph, where $\mathcal{W}$ denotes the set of nodes and $\mathcal{E}$ is the set of edges. Specifically, $\mathcal{W}$ consists of three types of nodes, which represents the three types of entities: users $u_i \in \mathcal{U}$ with $i \in \{1, \ldots, N_u\}$, hashtags $h_j \in \mathcal{H}$ with $j \in \{1, \ldots, N_h\}$, micro-videos $v_k \in \mathcal{V}$ with $k \in \{1, \ldots, N_v\}$, and $\mathcal{U} \cup \mathcal{H} \cup \mathcal{V} = \mathcal{W}$. Whereinto, $\mathbf{V} \in \mathbb{R}^{N_v \times D}$ is a feature matrix with $\mathbf{v}_k \in \mathbb{R}^{D_v}$ representing the feature vector of micro-video node v_k. D_v is the length of the feature vector. For the ease of presentation, i, j, and k will be assigned to index user, hashtag, and micro-video, respectively. When an interaction exists between two nodes w_i and w_k (e.g, a user u_i uploads a micro-video v_k), there will be an edge $e_{ik} = (w_i, w_k) \in \mathcal{E}$ to link the two nodes in the graph. In our task, we only consider the interactions among different types of nodes, namely, user-hashtag, hashtag-microvideo, and microvideo-user. Let $\mathcal{H}_i$ denote the hashtag neighbors of user u_i and $\mathcal{V}_i$ denote the micro-video neighbors of user u_i, namely, $\mathcal{H}_i$ is the hashtags set that u_i used to tag her uploaded micro-video set $\mathcal{V}_i$.

3.1.2 Model overview. When users provide hashtags to describe their uploaded micro-videos, they will select the hashtags *which (in their minds) can be used to describe the micro-video contents that interest them.* It implies that for the same micro-video, different users may be interested in different content and even two users are interested in the same content of a micro-video, it is still possible that they will use different hashtags due to their own preferences on hashtags. In other words, for each pair of <micro-video, hashtag>, the suitability score of the hashtag to the micro-video depends on the interests of the target user in the micro-video and personal opinions on the hashtag. In light of this, we propose a Graph Convolution Network based Personalized Hashtag Recommendation model **(GCN-PHR** for short). Given a micro-video v_k uploaded by the user u_i, for each candidate hashtag $h_j \in \mathcal{H}$, our model will (1) generate the representations $\bar{\mathbf{v}}_k^i$ and $\bar{\mathbf{h}}_j^i$ for micro-video v_k and the hashtag h_j respectively based on this user u_i's preferences, and then (2) compute the suitability score (or similarity score) of h_j with respect to v_k based on $\bar{\mathbf{v}}_k^i$ and $\bar{\mathbf{h}}_j^i$. Therefore, the core of our model is to learn the user-specific micro-video representation $\bar{\mathbf{v}}_k^i$ and user-specific hashtag representation $\bar{\mathbf{h}}_j^i$.

3.2 GCN-based Representation Learning

Intuitively, learning *a user-specific micro-video representation* $\bar{\mathbf{v}}_k^i$ needs to model the interaction between *the user preference on micro-videos* $\mathbf{u_i^v}$ and *the target micro-video* representation $\mathbf{v_k}$; similarly, learning *a user-specific hashtag representation* $\bar{\mathbf{h}}_j^i$ is to model the interaction between *the user preference on hashtags* $\mathbf{u_i^h}$ and *the target hashtag* representation $\mathbf{h_j}$. Thinking one step further, the hashtags that a user used to tag micro-videos actually reflect her preference on the micro-videos to some extent; and in turn, the micro-videos that a user tagged also contain her preference on the hashtags. In other words, $\mathcal{H}_i$ and $\mathcal{V}_i$ mutually represent each other and also reflect the user preference on each other. Therefore, it is beneficial to consider $\mathbf{u_i^h}$ into the modeling of $\bar{\mathbf{v}}_k^i$; and vice versa. In light of this, the user preference $\mathbf{u}_i$ is modeled based on both $\mathbf{u_i^v}$ and $\mathbf{u_i^h}$, and then used to learn $\bar{\mathbf{v}}_k^i$ and $\bar{\mathbf{h}}_j^i$.

Specifically, on the constructed graph the model, we adopt the message-passing idea [3] to learn the user preference on hashtags $\mathbf{u_i^h}$ and micro-videos $\mathbf{u_i^v}$ based on the user u_i's hashtag neighbors $\mathcal{H}_i$ and micro-video neighbors $\mathcal{V}_i$, respectively. After that, $\mathbf{u_i^h}$ and $\mathbf{u_i^v}$ are aggregated to learn the user preference $\mathbf{u_i}$. Finally, the $\bar{\mathbf{v}}_k^i$ is learned based on $\mathbf{u_i}$ and $\mathbf{v}_k$, and $\bar{\mathbf{h}}_j^i$ is learned based on $\mathbf{u_i}$ and $\mathbf{h}_j$. In the following, we introduce the learning of those representations sequentially.

3.2.1 User preference on hashtags. In our model, a user u_i preference on hashtags $\mathbf{u_i^h}$ is modeled by accumulating the incoming messages from all the neighbor hashtags $\mathcal{H}_i$. According to the idea of message-passing, the message transferred from a hashtag $h_j \in \mathcal{H}_i$ to the user u_i is defined as:

$$\mathbf{m}_{h_j \to u_i} = \mathbf{W}_h^u \mathbf{h}_j, \tag{1}$$

where $\mathbf{m}_{h_j \to u_i}$ denote the message vector from hashtag h_j to user u_i, and $\mathbf{W}_h^u$ is the weight matrix which maps the hashtag vector into the user embedding space. Based on this, $\mathbf{u_i^h}$ is defined as:

$$\mathbf{u}_i^h = \phi\Big(\frac{1}{|\mathcal{H}_i|} \sum_{h_j \in \mathcal{H}_i} \mathbf{m}_{h_j \to u_i}\Big), \tag{2}$$

$\phi(\cdot)$ is the activation function and $|\mathcal{H}_i|$ denotes the number of neighbor hashtags.

3.2.2 User preference on micro-videos. The user preference on micro-videos $\mathbf{u}_i^{\mathrm{v}}$ can be learned in the same way, namely, by accumulating the messages form all the neighbor videos $\mathcal{V}_i$. The passed message $\mathbf{v}_k$ from a micro-video represents all the contents in the video v_k. Notice that a micro-video contains a sequence of video frames with rich information. For a specific micro-video, a user may be only interested in its certain parts. To accurately model the user preference on micro-videos, it is crucial to identify which part in each micro-video attracts the attention of the user. Fortunately, the hashtags of a micro-video are usually provided based on the content interests the user in this micro-video. Therefore, the hashtags can better characterize the user preference on micro-videos. In light of this, in our model, the message from each video v_k to a user u_i is determined by its hashtags provided by the user. Given a hashtag h_j of a video v_k, we estimate its similarity by,

$$s_{jk} = g(\mathbf{h}_j, \mathbf{W}_v^h \mathbf{v}_k), \tag{3}$$

where $\mathbf{W}_v^h$ is a weight matrix and $g(\cdot)$ is a similarity function to measure the similarity of vectors. Different functions can be applied here (e.g., cosine function) and we use a fully connected layer to implement the function.

A hashtag can be used for many micro-videos by the user. Let $\mathcal{V}_{i,j}$ be the set of micro-videos tagged with the hashtag h_j by user u_i. we normalize the similarity score s_{jk} to obtain the relative similarity between a video and a hashtag based on the user preference.

$$\alpha_{jk} = \frac{exp(s_{jk})}{\sum_{v_{k'} \in \mathcal{V}_{i,j}} exp(s_{jk'})}, \tag{4}$$

α_{jk} is the normalized similarity score. Let $\mathcal{H}_{i,k}$ be the hashtag set of user u_i provided to micro-video v_k, the message of a video v_k passes to a user u_i is defined as:

$$\mathbf{m}_{v_k \to u_i} = \sum_{h_j \in \mathcal{H}_{i,k}} \alpha_{jk} \cdot \mathbf{W}_v^u \mathbf{v}_k, \tag{5}$$

It indicates that the message from v_k passes to the user u_i is dependent on v_k's aggregated similarity to all the hashtags that u_i provides.

Similar to Eq. 2, the user preference on micro-videos is the aggregation of the messages from all the neighbor micro-videos, namely,

$$\mathbf{u}_i^h = \phi\Big(\sum_{v_k \in \mathcal{V}_i} \mathbf{m}_{v_k \to u_i}\Big). \tag{6}$$

3.2.3 User representation learning. As discussed above, the user preference is obtained by combining the user preference on hashtags and on micro-videos. Many combination methods can be applied here, such as concatenation [9], addition [6], or more complicate deep fusion models [36]. In this work, we try two methods: network-based fusion and transformation-based summation.

Neural network-based fusion. In this method, $\mathbf{u}_i^v$ and $\mathbf{u}_i^h$ are first concatenated and then fed into a fully connected layer to obtain the final representation of the user preference. Formally, the user preference is obtained by,

$$\mathbf{u}_i = \phi(\mathbf{W}_{nn}\,[\mathbf{u}_i^v\ ,\ \mathbf{u}_i^h] + \mathbf{b}_{nn}), \tag{7}$$

where $[\cdot,\cdot]$ is the concatenation operator; $\mathbf{W}_{nn}$ and $\mathbf{b}_{nn}$ indicate the learnable weight matrix and bias vector in the fully connected layer, respectively.

Transformation-based summation. In this method, $\mathbf{u}_i^v$ and $\mathbf{u}_i^h$ are fist transformed into the same space for element-wise summation. Formally, the user preference is obtained by,

$$\mathbf{u}_i = \mathbf{W}_u^v\,\mathbf{u}_i^v + \mathbf{W}_u^h\,\mathbf{u}_i^h \tag{8}$$

where $\mathbf{W}_u^v$ and $\mathbf{W}_u^h$ denote the transformation matrices.

3.2.4 Hashtag representation learning. The hashtag representation is learned analogously with user preference learning. Specifically, for a hashtag h_j, its representation is based on both the messages from all its neighbor users and the messages from all its neighbor micro-videos. The message passed from a user to a hashtag is computed in the same way as the message passed from a hashtag to a user, and the message passed from a micro-video to a hashtag is computed in a similar way as the message passed from a micro-video to a user. To avoid the duplication of presentation, here we skip the detailed steps of how to compute the hashtag representation in our model.

Further, according to recent work [38], we can easily utilize message from multiple-hop neighbors by recursively stacking multiple layers to enforce the representations.

3.2.5 User-specific micro-video representation. The micro-video representation $\mathbf{v}_k$ is directly extracted from the content of the micro-video v_i, which is a concatenation of its visual, acoustic, and textual features. Detailed information about how to extract those multimodal features is described in Section 4.1. Based on the user preference $\mathbf{u}_i$ and micro-video representation $\mathbf{v}_k$, the user-specific micro-video representation $\bar{\mathbf{v}}_k^i$ is obtained by:

$$\bar{\mathbf{v}}_k^i = \phi(\mathbf{W}^v\ \mathbf{v}_k\ + \mathbf{W}_u^v\ \mathbf{u}_i + \mathbf{b}^v), \tag{9}$$

where $\mathbf{W}^v$, $\mathbf{W}_u^v$ and $\mathbf{b}^v$ denote the weight matrices and bias vector in the fully connected layer.

3.2.6 User-specific hashtag representation. Analogously, the user-specific hashtag representation $\bar{\mathbf{h}}_j^i$ is also learned via a fully connected layer based on the user preference $\mathbf{u}_i$ and hashtag representation $\mathbf{h}_j$,

$$\bar{\mathbf{h}}_j^i = \phi(\mathbf{W}^h\ \mathbf{h}_j\ + \mathbf{W}_u^h\ \mathbf{u}_i + \mathbf{b}^h), \tag{10}$$

where $\mathbf{W}^v$, $\mathbf{W}_u^v$ and $\mathbf{b}^v$ are the weight matrices and bias vector in the fully connected layer.

3.2.7 Personalized Hashtag Recommendation. Given a new micro-video v_k uploaded by a user u_i, the hashtags in $\mathcal{H}$ could be recommended in the descending order of their similarity score with respect to v_k based on based on u_i's preference. Specifically, the similarity score is computed by the dot product of the user-specific hashtag representation $\bar{\mathbf{h}}_j^i$ and the user-specific micro-video representation $\bar{\mathbf{v}}_k^i$, namely, $(\bar{\mathbf{h}}_j^i)^T \bar{\mathbf{v}}_k^i$.

3.3 Pairwise-based Learning

Similar to the ranking-oriented recommendation algorithm [4, 18], we adopt the pairwise-based learning method for optimization. To perform the pairwise ranking, it need to constructs a triplet of one micro-video v_k, one positive hashtag h_j, and one negative hashtag

Table 1: Statistics of the evaluation dataset. (#Micro-videos, #Hashtags, and #Users denote the numbers of micro-videos, Hashtags, and users, respectively.)

Dataset	#Micro-videos	#Users	#Hashtags
YFCC100M	134,992	8,126	23,054
Instagram	48,888	2,303	12,194

Table 2: Feature summarization of three modalities. (Visual, Acoustic, and Textual denote the dimensions of visual, acoustic, and textual modalities, respectively.)

Dataset	Visual	Acoustic	Textual
YFCC100M	2,048	128	100
Instagram	2,048	128	100

h'_j, where h_j is a hashtag of v_k and h'_j is not. Let $\mathcal{R} = \{(v_i, h_j, h'_j)\}$ be the triplet sets for training. The objective function can be formulated as

$$\arg\min_{\theta} \sum_{(\mathbf{v}_k, \mathbf{h}_j, \mathbf{h}'_j) \in \mathcal{R}} -\ln \phi((\bar{\mathbf{h}}_j^i)^T \bar{\mathbf{v}}_k^i - (\bar{\mathbf{h}}_j'^i)^T \bar{\mathbf{v}}_k^i) + \lambda \|\Theta\|_2^2, \quad (11)$$

where λ and Θ represent the regularization weight and the parameters of the model, respectively.

4 EXPERIMENTS

In this section, we first present the experimental settings (i.e. datasets, baselines, evaluation protocols, and parameter settings), followed by answering the above three questions and end up with some visualization examples.

4.1 Experimental Settings

Datasets. We conducted experiments on a public dataset YFCC100M [34] and our collected Instagram dataset. The characteristics of these datasets are summarized in Table 1.

YFCC100M.[5] The YFCC100M dataset is the largest publicly accessible multimedia collection, containing the metadata of around 99.2 million photos and 0.8 million videos from Flickr. In our task, we focus on the personalized micro-video hashtag recommendation, whereas we crawled video dataset from its API. And we also collected their user profiles and the annotated hashtags. This eventually crawled a dataset of 134, 992 micro-videos, 8, 126 users, and 23, 054 hashtags. Following that, we extracted a rich set of features from textual, visual and acoustic modalities. Specially, we employed FFmpeg[6] to extract the keyframes from micro-videos, and then employed the ResNet50 [17] model pre-trained by Pytorch[7] to extract the visual features. Simultaneously, we separated audio tracks from micro-videos with FFmpeg, and adopted VGGish [19] to learn the acoustic deep learning features. In addition, we utilized Sentence2Vector [2] trained with twitter text dataset to extract the textual features from micro-video descriptions.

Instagram. To construct a micro-video dataset for evaluating our method, we crawled the micro-video associated with the description from Instagram. We randomly started with some users and iteratively crawled the list of users who followed those users. In such a way, we harvested about 1 million micro-videos, 10 thousand users, and 100 thousand hashtags. To evaluate our method, we filter out the micro-video without any hashtag annotated, and removed users whose all micro-videos are discarded accordingly. After removing the micro-videos and users, we obtained a micro-video dataset of 48, 888 micro-videos, 2, 303 users, and 12, 194 hashtags. Similarly, we perform a similar process to extract the features of micro-videos in the Instagram dataset. For further clarification, we summarized the multimodal features of two datasets in Table 2.

Baselines. To evaluate the effectiveness of our model, we compared our proposed method with several state-of-the-art baselines.

- **UTM** [11]. The method utilizes a 3-way gating to combine heterogeneous features into a learning framework where the model is conditioned on the user profile. This model learns a joint d-dimensional embedding space for posts and hashtags.
- **ConTagNet** [32]. This baseline proposes a convolutional neural network (CNN) based framework to integrate user profiles with posts' content for hashtag prediction. This model consists of two components: CNN part and neural network part. In the experiment, for the fairness, we leveraged the extracted multimodel information and fed them into the fully connected network with the corresponding users' information to estimate the hashtags.
- **CSMN** [10]. The method named as Context Sequence Memory Network (CSMN) which leverages the historical information as the prior knowledge of users' vocabularies or writing styles. Towards this end, the content information extracted from micro-videos and hashtags users annotated is used as the memory retaining the context information. Besides, previously generated hashtags can be appended into memory to capture long-term information. With the proposed CNN memory structure, hashtags which meet the context information can be predicted.
- **USHM** [34]. This approach develops a user-specific hashtag model that takes the hashtag usage patterns of a user into account, and adopts a three-way tensor model to learn user embeddings. This baseline trains the model on image-user-hashtag triplets that allow the model to learn patterns in the hashtag usage of particular users and to disambiguate the learning signal.

Evaluation Protocols and Parameter Settings. We randomly split the dataset into training, validation, and testing sets with 8:1:1 ratio as in [5, 7, 8], and created the training triples based on random negative sampling. We used precision, recall, and accuracy to evaluate the recommended results. To train our proposed model, we use the *leaky_relu* as the activation function in our model and randomly initialized model parameters with a Gaussian distribution, optimizing the model with stochastic gradient descent (SGD). We tested the batch size of $\{128, 256, 512\}$, the latent feature dimension of $\{32, 64, 128\}$, the learning rate of $\{0.0001, 0.0005, 0.001.0.005, 0.01\}$ and the regularizer of $\{0, 0.00001, 0.0001, 0.001, 0.01, 0.1\}$. In addition, we sample 1, 000 negative hashtags for each micro-video in the test process. As the findings are consistent across the dimensions of latent vectors, if not specified, we only show the result of D=64, a relatively large number that returns good accuracy.

4.2 Performances Comparison (RQ1)

The comparative results are summarized in Table 3. From this table, we have the following observations:

[5] https://multimediacommons.wordpress.com/yfcc100m-core-dataset/.
[6] https://ffmpeg.org/.
[7] https://pytorch.org/.

Table 3: Performance comparison between our model and the baselines.

Model	YFCC100M						Instagram					
	P@5	P@10	R@5	R@10	A@5	A@10	P@5	P@10	R@5	R@10	A@5	A@10
UTM	0.2469	0.1780	0.3053	0.4108	0.6860	0.7724	0.5193	0.4503	0.2305	0.3659	0.7823	0.8136
ConTagNet	0.3595	0.2441	0.4358	0.5426	0.7856	0.8395	0.4391	0.3863	0.1982	0.3136	0.74.44	0.8178
CSMN	0.3027	0.1897	0.4004	0.4428	0.6403	0.6634	0.5230	0.4691	0.2614	0.3826	0.8246	0.8770
USHM	0.3891	0.2711	0.4865	0.6158	0.8398	0.8877	0.6613	0.5821	0.3402	0.5134	0.9107	**0.9365**
GCN-PHR	**0.4004**	**0.3053**	**0.5125**	**0.6667**	**0.8403**	**0.9021**	**0.6847**	**0.6286**	**0.4075**	**0.5667**	**0.9134**	0.9332
%Improv.	**2.90%**	**12.61%**	**5.34%**	**8.30%**	**0.01%**	**1.62%**	**3.54%**	**7.99%**	**19.78%**	**10.38%**	**0.30%**	-0.35%

Table 4: Performance comparison between our model and the variants.

Model	YFCC100M						Instagram					
	P@5	P@10	R@5	R@10	A@5	A@10	P@5	P@10	R@5	R@10	A@5	A@10
Variant-I	0.3786	0.2745	0.4923	0.6266	0.8335	0.8901	0.6703	0.58330.	3473	0.5232	0.9068	0.9245
Variant-II	0.3829	0.2837	0.5031	0.6401	0.8307	0.8834	0.6765	0.6053	0.3735	0.5398	0.9089	0.9306
Variant-III	0.3801	0.2754	0.4966	0.6312	0.8283	0.8831	0.6725	0.5974	0.3665	0.5274	0.8953	0.9268
GCN-PHR	**0.4004**	**0.3053**	**0.5125**	**0.6667**	**0.8403**	**0.9021**	**0.6847**	**0.6286**	**0.4075**	**0.5667**	**0.9134**	**0.9332**

- Without a doubt, our proposed method achieves the best performance on three datasets. Especially, it outperforms the state-of-the-art baselines over the datasets, verifying the effectiveness of our model. Furthermore, it shows the rationality of the user preference and hashtag semantic modeling based on the micro-video content information.
- The indexes indicate that the USHM is better than the other baselines for it is the first to consider the hashtag personalized recommendation. It is beneficial for hashtag recommendation to capture the hashtag usage pattern of the users based on the interaction between user and hashtag. However, in comparison, this method is inferior to the one we proposed. Because the USHM merely models users' interactions between hashtags and posts, instead of profiling them.
- The precision of our method on YFCC grew faster than that on Instagram, while the increase in recall is just the opposite. These phenomenons result from the average hashtag number of every single micro-video is more than YFCC100M.
- In terms of accuracy, our method did not have significant improvement over USHM, and even lower in some cases. We suggest that the accuracy merely calculate the ratio of at least one ground truth hashtag appears in the top K, which results in the approximation of the results.
- Unexpectedly, the CSMN model shows the worst results on YFCC100M. The reason for this may be due to the CSMN is more susceptible to the density of the dataset. To predict the hashtag, CSMN leverages the previous hashtag to represent the user. On the sparse dataset, the bias is accumulated during the whole prediction and causes worse performance.

4.3 Model Study

In this section, we studied the effectiveness of each component of our proposed model. We listed the following several variants to compare with the proposed model.

- **Variant-I.** In this model, we built a user-hashtag-microvideo graph and performed graph convolution operations on each node without the attention mechanism. This variant is designed to investigate the effectiveness of attention mechanism in our model.
- **Variant-II.** To evaluate the effectiveness of the user preference to hashtag representation, we discard the user preference attended on the content information and passed the whole information from micro-video nodes to the hashtag nodes.
- **Variant-III.** In contrast to **Variant-II**, this variant removes the hashtag guiding the micro-video feature extraction and aggregates the micro-video content information straightway, to evaluate the effectiveness of hashtag to user preference modeling.

In Table 4, we have the following observations:

- As expected, GCN-PHR outperforms other variants, with more significant improvement on the Instagram dataset. This demonstrates the effectiveness of the user and hashtag representations. It is due to that the user preference not only affects the micro-video features extraction, but influences the hashtag semantic representations, which is the core of the hashtag personalized recommendation.
- Comparing **Variant-II** with **Variant-III**, we found that the former is better than the later one. We conjecture that the micro-video contains abundant information and it is hard to capture the information from the content to represent the hashtag without the user preference guiding.
- In terms of the A@5 and A@10, we observed that the methods have no significant results. This is because that the accuracy gives a measure of how often at least one of the ground truth hashtags appears in the top 5 (10) ranked hashtags. It is largely indifferent to the number of ground truth hashtags.
- We compared the three variants, and the result demonstrates that the user preference and hashtag representations are improved by content information. In addition, the user preference and hashtag semantic can help the feature extraction during the message-passing, facilitating the hashtag and user preference representations.

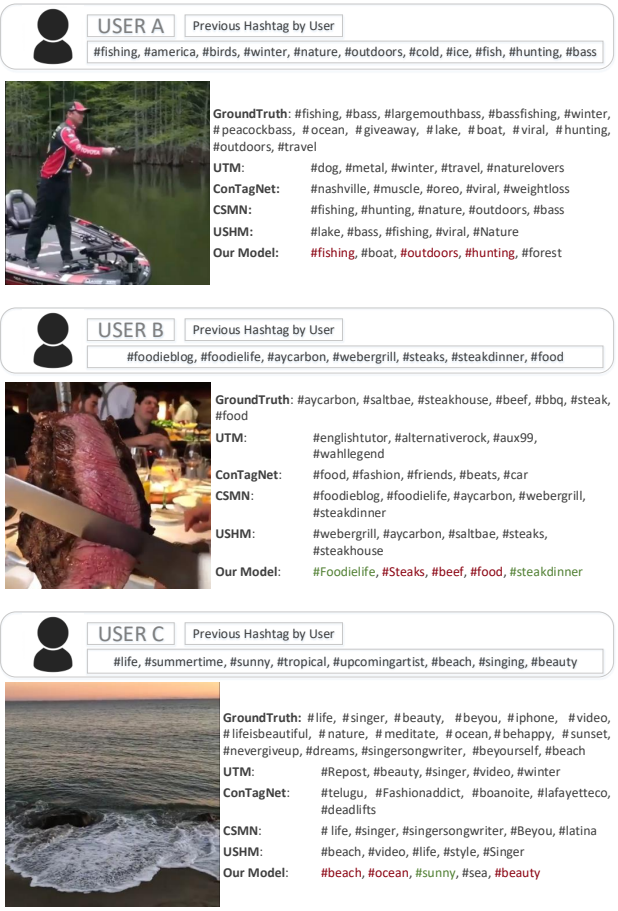

Figure 3: Visualization of hashtag recommended by baselines and our model on the Instagram dataset.

4.4 Visualization

To evaluate our method, we randomly selected a number of micro-videos for visualization testing, taking three specific micro-videos as examples. In the figure, three micro-videos from different users, provide the user historical hashtag usage and baseline prediction for comparison, so as to demonstrate our hypothesis of user hashtag usage pattern. Whereinto, the hashtag marked red is predicted accurately, and the tag marked green indicates that the hashtag has appeared in the historical information, but has not marked the current micro-video. According to the three examples, we gained conclusions as following:

- In the first example, it can be seen that our method can accurately recommend the hashtags of '#fishing', '#hunting' '#boat', and '#outdoors'. Whereinto, '#hunting', '#fishing', and '#outdoors' are all hashtags that users have used before, which shows that our method has captured the characteristics of user hashtag usage pattern. USHM and CSMN also recommend '#fishing', '#hunting', and '#outdoors' based on the correct usage habits of hashtag users. However, these methods ignore the '#boat'. This is because this hashtag does not appear in the user's historical hashtags, so it is difficult to recommend it merely based on modeling user's hashtag usage habits. Different from these methods, our method combines and filters the content information to express the hashtag, and accurately recommends the '#boat' to the user.
- As shown in the second example, our model predicts hashtags '#steaks', '#beef', and '#food' to the micro-video, especially, '#beef' has not been used by users. This exhibits that our method leverages the micro-video content information to represent the hashtag semantic, which facilitates the hashtag recommendation to micro-videos. Moreover, although the model mistakenly recommends '#foodielife' and '#steakdinner', we found that they are fit for the micro-video content and user's hashtag usage pattern. Therefore, it also demonstrates that our proposed model is reasonable.
- In the last example, our presented model recommends the third micro-video with '#beach', '#ocean', and '#beauty'. Especially, the '#beauty' hashtag is distant with these hashtags, since it is hard to describe with some fixed information and varying with the different users. This phenomenon verifies that our hashtag representation based on content information bridges the gap between the user's opinion and the visual, acoustic, and textual information to some degree.
- Above of all, we observed that our proposed method outperforms the other baselines, especially for the micro-videos containing more concepts. We believe that our proposed model can extract the features from micro-video content information with user preference and annotate the appropriate hashtags in accord with the user's hashtag usage pattern.

5 CONCLUSION

In a real-world scenario, when users provide hashtags to describe their uploaded micro-videos, they tend to select the hashtags to describe the micro-video contents that they are interested in. It implies that for the same micro-video, different users may be interested in different content and sometimes two users are interested in the same content of a micro-video, yet they may use different hashtags based on their personal preferences on hashtags. To model such complex correlations, we propose a Graph Convolution Network based Personalized Hashtag Recommendation model. Specifically given a micro-video uploaded by the user, for each candidate hashtag, our model (1) generate the representations for micro-video and the hashtag respectively based on the user preference, and then (2) compute the suitability score (or similarity score) of hashtag with respect to the micro-video based on the user preference. To demonstrate the effectiveness of our model, we comparatively justified it over two micro-video benchmark datasets. The experimental results show that our proposed model outperforms several state-of-the-art baselines. In addition, we randomly sample some micro-videos and visualize the results of our proposed method and several baselines to verify the performance between the methods.

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] Hamidreza Alvari. 2017. Twitter hashtag recommendation using matrix factorization. *arXiv preprint arXiv:1705.10453* (2017).

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of International Conference of Learning Representation*.

[3] Rianne van den Berg, Thomas N Kipf, and Max Welling. 2017. Graph convolutional matrix completion. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[4] Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences. In *The World Wide Web Conference*. ACM, 151–161.

[5] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems* 37, 2 (2019), 16.

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

[7] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of World Wide Web Conference*. 639–648.

[8] Zhiyong Cheng, Shen Jialie, and Steven CH Hoi. 2016. On effective personalized music retrieval by exploring online user behaviors. In *Proceedings of International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 125–134.

[9] Zhiyong Cheng and Jialie Shen. 2016. On effective location-aware music recommendation. *ACM Transactions on Information Systems* 34, 2 (2016), 13.

[10] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 895–903.

[11] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1731–1740.

[12] Kuntal Dey, Ritvik Shrivastava, Saroj Kaushik, and L Venkata Subramaniam. 2017. Emtagger: a word embedding based novel method for hashtag recommendation on twitter. In *Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1025–1032.

[13] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. Protein interface prediction using graph convolutional networks. In *Advances in Neural Information Processing Systems*. 6530–6539.

[14] Victor Garcia and Joan Bruna. 2018. Few-Shot Learning with Graph Neural Networks. In *Proceedings of International Conference on Learning Representations*. 1–12.

[15] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. (2017).

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *Proceedings of International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 355–364.

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proceedings of IEEE international conference on acoustics, speech and signal processing*. IEEE, 131–135.

[20] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1–10.

[21] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations*. 1–14.

[22] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. 2017. Distance Metric Learning Using Graph Convolutional Networks: Application to Functional Brain Networks. In *International Conference on Medical Image Computing & Computer-assisted Intervention*. 469–477.

[23] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. 2018. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision*. 335–351.

[24] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. 2018. Learning deep generative models of graphs. In *Proceedings of the International Conference on Machine Learning*. 1–22.

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

[26] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[27] Kenneth Marino, Ruslan Salakhutdinov, and Harikrishna Mulam. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 20–28.

[28] Federico Monti, Michael Bronstein, and Xavier Bresson. 2017. Geometric matrix completion with recurrent multi-graph neural networks. In *Advances in Neural Information Processing Systems*. 3697–3707.

[29] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian. 2017. Enhancing Micro-video Understanding by Harnessing External Sounds. In *Proceedings of ACM Multimedia Conference on Multimedia Conference*. 1192–1200.

[30] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *Proceedings of the International conference on machine learning*. 2014–2023.

[31] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero Moreno, Ben Glocker, and Daniel Rueckert. 2017. Spectral graph convolutions for population-based disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 177–185.

[32] Yogesh Singh Rawat and Mohan S Kankanhalli. 2016. ConTagNet: Exploiting user context for image tag recommendation. In *Proceedings of the ACM international conference on Multimedia*. ACM, 1102–1106.

[33] Van Cuong Tran, Dosam Hwang, and Ngoc Thanh Nguyen. 2018. Hashtag Recommendation Approach Based on Content and User Characteristics. *Cybernetics and Systems* 49, 5-6 (2018), 368–383.

[34] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. 2018. Separating self-expression and visual content in hashtag supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5919–5927.

[35] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia* 14, 4 (2012), 975–985.

[36] Meng Wang, Changzhi Luo, Bingbing Ni, Jun Yuan, Jianfeng Wang, and Shuicheng Yan. 2017. First-person daily activity recognition with manipulated object proposals and non-linear feature fusion. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 2946–2955.

[37] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining*.

[38] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174.

[39] 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 the ACM international conference on Information and knowledge management*. ACM, 1031–1040.

[40] Yilin Wang, Suhang Wang, Jiliang Tang, Guojun 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.

[41] Zhenghua Xu, Thomas Lukasiewicz, Cheng Chen, Yishu Miao, and Xiangwu Meng. 2017. Tag-Aware Personalized Recommendation Using a Hybrid Deep Model. In *Proceedings of International Joint Conference on Artificial Intelligence*. 3196–3202.

[42] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 974–983.

[43] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*. 6412–6422.

[44] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2018. Learning Multimodal Taxonomy via Variational Deep Graph Embedding and Clustering. In *Proceedings of the ACM international conference on Multimedia*. 681–689.

[45] Qi Zhang, Yeyun Gong, Xuyang Sun, and Xuanjing Huang. 2014. Time-aware personalized hashtag recommendation on social media. In *Proceedings of International Conference on Computational Linguistics: Technical Papers*. 203–212.

[46] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, 13 (2018), i457–i466.