

# MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video

Yinwei Wei  
Shandong University  
weiyinwei@hotmail.com

Xiang Wang\*  
National University of Singapore  
xiangwang@u.nus.edu

Liqiang Nie  
Shandong University  
nieliqiang@gmail.com

Xiangnan He  
University of Science and Technology  
of China  
xiangnan@comp.nus.edu.sg

Richang Hong  
Hefei University of Technology  
hongrc@hfut.edu.cn

Tat-Seng Chua  
National University of Singapore  
chuats@comp.nus.edu.sg

## ABSTRACT

Personalized recommendation plays a central role in many online content sharing platforms. To provide quality micro-video recommendation service, it is of crucial importance to consider the interactions between users and items (*i.e.*, micro-videos) as well as item contents from various modalities (*e.g.*, visual, acoustic, and textual). Existing works on multimedia recommendation largely exploit multi-modal contents to enrich item representations, while less effort is made to leverage information interchange between users and items to enhance user representations and further capture user's fine-grained preferences on different modalities.

In this paper, we propose to exploit user-item interactions to guide the representation learning in each modality, and further personalized micro-video recommendation. We design a multi-modal Graph Convolution Network (MMGCN) framework built upon the message-passing idea of graph neural networks, which can yield modal-specific representations of users and micro-videos to better capture user preferences. Specifically, we construct a user-item bipartite graph in each modality, and enrich the representation of each node with the topological structure and features of its neighbors. Through extensive experiments on three public datasets: Tiktok, Kwai, and MovieLens, we demonstrate that our proposed model significantly outperforms state-of-the-art multi-modal recommendation methods.

## CCS CONCEPTS

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

## KEYWORDS

Graph Convolution Network, Multi-modal Recommendation, Micro-video Understanding

\*Xiang Wang and Liqiang Nie are the corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

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

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6889-6/19/10...\$15.00

<https://doi.org/10.1145/>

## ACM Reference Format:

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings 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

Personalized recommendation has become a core component in many online content sharing services, spanning from image, blog to music recommendation. Recent success of micro-video sharing platforms, such as Tiktok and Kwai, bring increasing attentions to micro-video recommendation. Distinct from these item contents (*e.g.*, image, music) that are solely from a single modality, micro-videos contain rich multimedia information — frames, sound tracks, and descriptions — and further involve multiple modalities — the visual, acoustic, and textual ones [24, 25, 27].

Incorporating such multi-modal information into historical interactions between users and micro-videos help establish an in-depth understanding of user preferences:

- There is a semantic gap between different modalities. Taking Figure 1 as an example, while having visually similar frames, micro-videos  $i_1$  and  $i_2$  have dissimilar textural representations due to different topic words. In such cases, ignoring such modality difference would mislead modeling of item representations.
- A user might have different tastes on modalities of a micro-video. For example, a user is attracted by the frames, but might turn out to be disappointed with its poor sound tracks. Multiple modalities, hence, have varying contributions to user preferences.
- Different modalities serve as different channels to explore user interests. In Figure 1, if user  $u_1$  cares more about frames,  $i_2$  is more suitable to be recommended; whereas,  $u_1$  might click  $i_3$  due to the interested textural descriptions.

Therefore, it is of crucial importance to distinguish and consider modal-specific user preferences.

However, existing works on multimedia recommendation [8, 17] mainly treat multi-modal information as a whole and incorporate them into a collaborative filtering (CF) framework, while lacking modeling of modal-specific user preferences. Specifically, multi-modal features of each item are unified as a single representation, reflecting their content similarity; thereafter, such representations are incorporated with user and item representations derived from CF framework, such as MF [29]. For instance, VBPR [17]

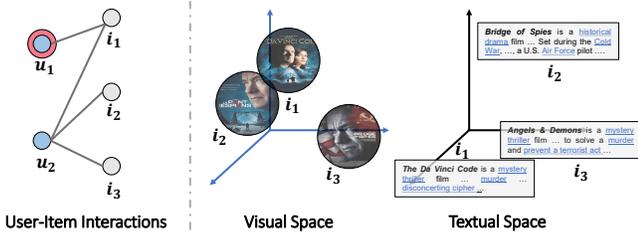


Figure 1: An illustration of modal-specific user preferences.

leverages visual features to enrich ID embeddings of items; ACF [8] employs the attention mechanism on a user’s history to encode two-level personal tastes on historical items and item contents into user representations. Such signals can be summarized as the paths connecting the target user and item based on historical interactions [38, 41]. For example,  $p_1 := u_1 \rightarrow i_1 \rightarrow u_2 \rightarrow i_2$  and  $p_2 := u_1 \rightarrow i_1 \rightarrow u_2 \rightarrow i_3$  suggest  $i_2$  and  $i_3$  are likely to be of interest to  $u_1$ . However, we argue that these signals are not sufficient to yield satisfactory performance. The key reason is that they ignore differences among modalities, further lacks an explicit modeling of modal-specific preferences of users. For example, in the visual modality,  $p_1$  suggests  $i_2$  is more likely to be adopted than  $i_3$  in  $p_2$ ; whereas,  $p_2$  presents more positive signals for recommending  $i_3$  in the textual modality.

To address the limitations, we focus on information interchange between users and items in multiple modalities. Inspired by the recent success of graph convolution networks (GCNs) [14, 22], we use the information-propagation mechanism to encode high-order connectivity between users and micro-videos in each modality, so as to capture user preference on modal-specific contents. Towards this end, we propose a *Multi-modal Graph Convolution Network* (MMGCN). Specifically, we construct a user-item bipartite graph on each modality. Intuitively, the historical behaviors of users reflect personal interests; meanwhile, the user groups can also profile items [38, 41]. Hence, in each modality (e.g., visual), we aggregate signals from the corresponding contents (e.g., frames) of interacted items and incorporate them into user representations; meanwhile, we boost the representation of an item with its user group. By performing such aggregation and combination operators recursively, we can enforce the user and item representations to capture signals from multi-hop neighbors, such that a user’s modal-specific preference is represented well in her representation. Ultimately, the prediction of a unseen interaction can be calculated as similarities between the user and micro-video representations. To demonstrate our method, we validate the framework over three publicly accessible datasets — Tiktok, Kwai, and Movielens. Experimental results show that our model can yield promising performance. Furthermore, we visualized user preference on different modalities, which shows the inconsistency of modal-specific preferences.

The main contributions of this work are threefold:

- We explore how information interchange on various modalities reflects user preferences and affects recommendation performance.
- We develop a new method MMGCN, which employs information propagation on the modality-aware bipartite user-item graph,

to obtain better user representations based on item content information.

- Extensive experiments are performed on three public datasets to demonstrate that our proposed model outperforms several state-of-the-art recommendation methods. In addition, we released our codes, parameters, and involved baselines to facilitate other researchers<sup>1</sup>.

## 2 MODEL FRAMEWORK

In this section, we elaborate our framework. As illustrated in Figure 2, our framework consists of three components — aggregation layers, combination layers, and prediction layer. By stacking multiple aggregation and combination layers, we encode the information interchange of users and items into the representation learning in each modality. Lastly, we fuse multi-modal representations to predict the interaction between each user and each micro-video in the prediction layer. In what follows, we detail each component.

### 2.1 Modality-aware User-Item Graphs

Instead of unifying multi-modal information, we treat each modality individually. Particularly, we have historical interactions (e.g., view, browse, or click) between users and micro-videos. Here we represent the interaction data as a bipartite user-item graph  $\mathcal{G} = \{(u, i) | u \in \mathcal{U}, i \in \mathcal{I}\}$ , where  $\mathcal{U}$  and  $\mathcal{I}$  separately denote the user and micro-video sets. An edge  $y_{ui} = 1$  indicates an observed interaction between user  $u$  and micro-video  $i$ ; otherwise  $y_{ui} = 0$ .

Beyond the interactions, we have multiple modalities for each micro-video — visual, acoustic, and textual features. For simplicity, we use  $m \in \mathcal{M} = \{v, a, t\}$  as the modality indicator, where  $v$ ,  $a$ , and  $t$  represent the visual, acoustic, and textual modalities, respectively. To accurately capture the users’ preferences on a particular modality  $m$ , we split the bipartite graph  $\mathcal{G}_m$  from  $\mathcal{G}$  by keeping the features in modality  $m$  solely.

### 2.2 Aggregation Layer

Intuitively, we can utilize the interaction data to enrich the representations of users and items. To be more specific, historical interactions of a user can describe her interest and capture the behavior similarity with others; meanwhile, the user group of a micro-video can provide complementary data to its multi-modal contents. We hence incorporate the information interchange into the representation learning.

Inspired by the message-passing mechanism of GCN, for a user (or micro-video) node in the bipartite graph  $\mathcal{G}_m$ , we employ an aggregation function  $f(\cdot)$  to quantify the influence (i.e., the representation being propagated) from its neighbors and output a representation as follows:

$$\mathbf{h}_m = f(\mathcal{N}_u), \quad (1)$$

where  $\mathcal{N}_u = \{j | (u, j) \in \mathcal{G}_m\}$  denotes the neighbors of user  $u$ , i.e., interacted micro-videos. We implement  $f(\cdot)$  via:

- **Mean Aggregation** employs the average pooling operation on the modal-specific features, and applies a nonlinear

<sup>1</sup><https://github.com/weiyinwei/MMGCN>.

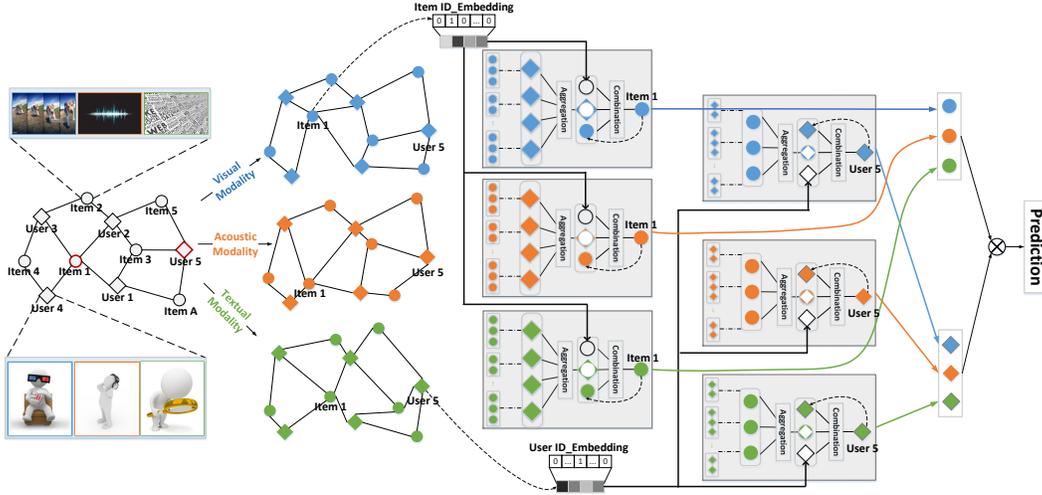


Figure 2: Schematic illustration of our proposed MMGCN model. It constructs the user-microvideo bipartite graph for each modality to capture the modal-specific user preference for the personalized recommendation of micro-video.

transformation, as follows:

$$f_{\text{avg}}(\mathcal{N}_u) = \text{LeakyReLU}\left(\frac{1}{|\mathcal{N}_u|} \sum_{j \in \mathcal{N}_u} \mathbf{W}_{1,m} \mathbf{j}_m\right), \quad (2)$$

where  $\mathbf{j}_m \in \mathbb{R}^{d_m}$  is the  $d_m$ -dimension representation of micro-video  $j$  in modality  $m$ ;  $\mathbf{W}_{1,m} \in \mathbb{R}^{d'_m \times d_m}$  is the trainable transformation matrix to distill useful knowledge, where  $d'_m$  is the transformation size; and we select  $\text{LeakyReLU}(\cdot)$  as the nonlinear activation function [38, 41]. Such aggregation method assumes that different neighbors would have the same contributions to the representation of user  $u$ , namely, user  $u$  is influenced equally by her neighbors.

- **Max Aggregation** leverages the max pooling operation to perform dimension-aware feature selection, as follows:

$$f_{\text{max}}(\mathcal{N}_u) = \text{LeakyReLU}\left(\max_{j \in \mathcal{N}_u} \mathbf{W}_{1,m} \mathbf{j}_m\right), \quad (3)$$

where each dimension of  $\mathbf{h}_m$  is set as that of the corresponding neighbor value with the maximum value. As such, different neighbors have varying contributions to the output representations.

As such, the aggregation layer is capable of encoding the structural information and distribution of neighbors into the representation of the ego user; analogously, we can update representations for item nodes.

### 2.3 Combination Layer

While containing the information being propagated from neighbors, such representations forgo user  $u$ 's own feature and the interaction among different modalities. However, existing GNN efforts (e.g., GCN [22], GraphSage [14], GAT [32]) only consider homogeneous features from one data source. Hence, directly applying their combination operations fails to capture the interactions between different modalities.

In this section, we present a new combination layer, which integrates the structural information  $\mathbf{h}_m$ , the intrinsic information  $\mathbf{u}_m$ , and the modality connection  $\mathbf{u}_{id}$  into a unified representation,

which is formulated as:

$$\mathbf{u}_m^{(1)} = g(\mathbf{h}_m, \mathbf{u}_m, \mathbf{u}_{id}), \quad (4)$$

where  $\mathbf{u}_m \in \mathbb{R}^{d_m}$  is the representation of user  $u$  in modality  $m$ ;  $\mathbf{u}_{id} \in \mathbb{R}^d$  is the  $d$ -dimension embedding of user ID, remaining invariant and serving as the connection across modalities.

Inspired by prior work [3] on multi-modal representation, we first apply the idea of the coordinated fashion, namely, separately project  $\mathbf{u}_m, \forall m \in \mathcal{M}$  into the latent space that is the same as  $\mathbf{u}_{id}$ :

$$\hat{\mathbf{u}}_m = \text{LeakyReLU}(\mathbf{W}_{2,m} \mathbf{u}_m) + \mathbf{u}_{id}, \quad (5)$$

where  $\mathbf{W}_{2,m} \in \mathbb{R}^{d \times d_m}$  is the trainable weight matrix to transfer  $\mathbf{u}_m$  into the ID embedding space. As such, the representations from different modalities are comparable in the same hyperplane; meanwhile, the ID embedding  $\mathbf{u}_{id}$  essentially bridges the gap between modal-specific representations, and propagates information across modalities during the gradient back-propagation process. In this work, we implement the combination function  $g(\cdot)$  via the following two methods:

- **Concatenation Combination** concatenates two representations, followed by a nonlinear transformation:

$$g_{\text{co}}(\mathbf{h}_m, \mathbf{u}_m, \mathbf{u}_{id}) = \text{LeakyReLU}\left(\mathbf{W}_{3,m}(\mathbf{h}_m || \hat{\mathbf{u}}_m)\right), \quad (6)$$

where  $||$  is the concatenation operation,  $\mathbf{W}_{3,m} \in \mathbb{R}^{d'_m \times (d'_m + d)}$  is the trainable model parameters.

- **Element-wise Combination** considers the element-wise feature interaction between two representations:

$$g_{\text{ele}}(\mathbf{h}_m, \mathbf{u}_m, \mathbf{u}_{id}) = \text{LeakyReLU}\left(\mathbf{W}_{3,m} \mathbf{h}_m + \hat{\mathbf{u}}_m\right), \quad (7)$$

where  $\mathbf{W}_{3,m} \in \mathbb{R}^{d \times d'_m}$  denotes a weight matrix to transfer the current representations into the common space. As such, the element-wise interactions between two features is taken into consideration, while two representations are assumed to be independent in Concatenation Combination.

**Table 1: Statistics of the evaluation dataset. (V, A, and T denote the dimensions of visual, acoustic, and textual modalities, respectively.)**

Dataset	#Interactions	#Items	#Users	Sparsity	V	A	T
Tiktok	726,065	76,085	36,656	99.99%	128	128	128
Kwai	1,664,305	329,510	22,611	99.98%	2,048	-	128
MovieLens	1,239,508	5,986	55,485	99.63%	2,048	128	100

## 2.4 Model Prediction

By stacking more aggregation and combination layers, we explore the higher-order connectivity inherent in the user-item graphs. As such, we can gather the information propagated from the  $l$ -hop neighbors in modality  $m$ , mimicking the exploration process of users. Formally, the representation from  $l$ -hop neighbors of user  $u$  and the output of  $l$ -th multi-modal combination layer are recursively formulated as:

$$\mathbf{h}_m^{(l)} = f(\mathcal{N}_u) \quad \text{and} \quad \mathbf{u}_m^{(l)} = g(\mathbf{h}_m^{(l)}, \mathbf{u}_m^{(l-1)}, \mathbf{u}_{id}), \quad (8)$$

where  $\mathbf{u}_m^{(l-1)}$  is the representation generated from the previous layer, memorizing the information from her  $(l-1)$ -hop neighbors.  $\mathbf{u}_m^{(0)}$  is set as  $\mathbf{u}_m$  at the initial iteration. Wherein, user  $u$  is associated with trainable vectors  $\mathbf{u}_m, \forall m \in \mathcal{M}$ , which are randomly initialized; whereas, item  $i$  is associated with pre-extracted features  $\mathbf{i}_m, \forall m \in \mathcal{M}$ . As a result,  $\mathbf{u}_m^{(l-1)}$  characterizes the user preferences on item features in modality  $m$ , and considers the influence of modality interactions that reflect the underlying relationships between modalities.

After stacking  $L$  single-modal aggregation and multi-modal combination layers, we obtain the final representations for user  $u$  and micro-video  $i$  via the linear combination of multi-modal representations, as:

$$\mathbf{u}^* = \sum_{m \in \mathcal{M}} \mathbf{u}_m^{(L)} \quad \text{and} \quad \mathbf{i}^* = \sum_{m \in \mathcal{M}} \mathbf{i}_m^{(L)} \quad (9)$$

## 2.5 Optimization

To predict the interaction between the users and micro-videos, we fuse their modal-specific representations of them and apply Bayesian Personalized Ranking (BPR) [29], which is a well-known pairwise ranking optimization framework, as the learning model. In particular, we model a triplet of one user and two micro-videos, in which one of the micro-videos is observed and the other one is not, formally as,

$$\mathcal{R} = \{(u, i, i') | (u, i) \in \mathcal{G}, (u, i') \notin \mathcal{G}\}, \quad (10)$$

where  $\mathcal{N}(u)$  consists of all micro-videos associated with  $u$ , and  $\mathcal{R}$  is a set of triples for training. Further, it is assumed that the user prefers the observed micro-video rather than the unobserved one. Then, the objective function can be formulated as,

$$\mathcal{L} = \sum_{(u, i, i') \in \mathcal{R}} -\ln \mu(\mathbf{u}^{*\top} \mathbf{i}^* - \mathbf{u}^{*\top} \mathbf{i}'^*) + \lambda \|\Theta\|_2^2, \quad (11)$$

where  $\mu(\cdot)$  is the sigmoid function;  $\lambda$  and  $\Theta$  represent the regularization weight and the parameters of the model, respectively.

## 3 EXPERIMENTS

In this section, we conduct experiments on three public datasets, aiming to answer the following research questions:

- **RQ1:** How does MMGCN perform compared with state-of-the-art multi-modal recommendation systems and other GNN-based methods on our task?
- **RQ2:** How do different designs (e.g., number of modalities, number of layers, selection of combination layer) influence the performance of MMGCN?
- **RQ3:** Can MMGCN capture the inconsistent preference of users on different modalities?

In what follows, we first present the experimental settings (i.e., datasets, baselines, evaluation protocols, and parameter settings), followed by answering the above three questions.

### 3.1 Experimental Settings

**Datasets.** To evaluate our model, we experimented with three public datasets: Tiktok, Kwai, and MovieLens. The characteristics of these datasets are summarized in Table 1.

- **Tiktok**<sup>2</sup>: The first dataset is published by Tiktok, which is a micro-video sharing platform that allows users to create and share micro-videos with duration of 3-15 seconds. It consists of users, micro-videos and their interactions (e.g., click). The micro-video features in each modality are extracted and published without providing the raw data. Particularly, the textual features are extracted from the micro-video captions given by users.
- **Kwai**<sup>3</sup>: As a popular micro-video service provider, Kwai has constructed a large-scale micro-video dataset. Similar with the Tiktok dataset, this contains the privacy-preserving user information, content features of micro-videos, and interaction data. However, the acoustic information of micro-videos is missing.
- **MovieLens**<sup>4</sup>: This movie dataset has been widely used to evaluate recommendations. To construct a dataset, we collected the titles and descriptions of movies from the MoiveLens-10M dataset and crawled the corresponding trailers instead of the full-length videos from Youtube<sup>5</sup>. We use the pre-trained ResNet50 [16] models to extract the visual features from key frames extracted from micro-video. In terms of acoustic modality, we separate audio tracks with FFmpeg<sup>6</sup> and adopt VGGish [20] to learn the acoustic deep learning features. For textual modality, we use Sentence2Vector [1] to derive the textual features from micro-videos' descriptions.

**Baselines.** To evaluate the effectiveness of our model, we compare MMGCN with the following state-of-the-art baselines, which can be grouped into two categories: CF-based (VBPR and ACF) and GCN-based (GCMC and GraphSAGE) methods.

- **VBPR** [17]. Such model integrates the content features and ID embeddings of each item as its representation, and uses the matrix factorization (MF) framework to reconstruct the historical interactions between users and items. In the experiments, we use

<sup>2</sup><http://ai-lab-challenge.bytedance.com/tce/vc/>.

<sup>3</sup><https://www.kuaishou.com/activity/uimc>.

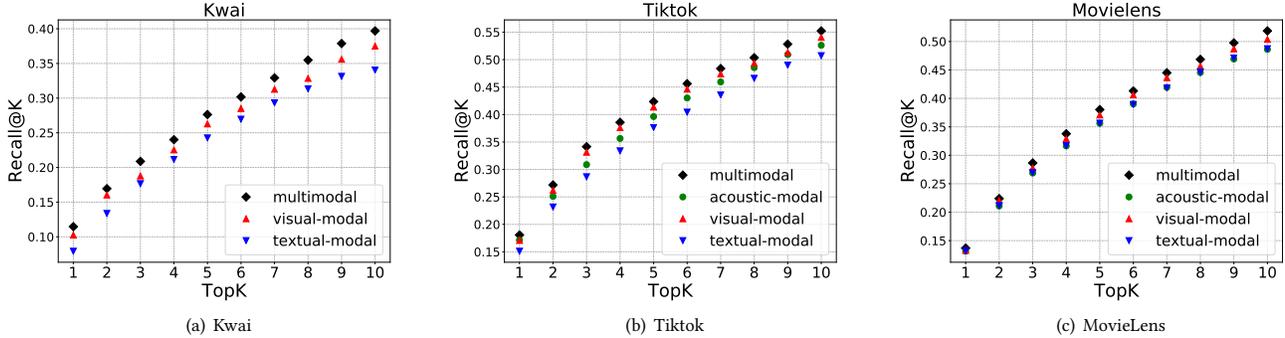
<sup>4</sup><https://grouplens.org/datasets/movielens/>.

<sup>5</sup><https://www.youtube.com/>.

<sup>6</sup><http://ffmpeg.org/>.

**Table 2: Performance comparison between our model and the baselines.**

Model	Kwai			Tiktok			MovieLens		
	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG
VBPR	0.2673	0.3386	0.1988	0.0972	0.4878	0.3136	<b>0.1172</b>	<b>0.4724</b>	<b>0.2852</b>
ACF	0.2559	0.3248	0.1874	0.8734	0.4429	0.2867	0.1078	0.4304	0.2589
GraphSAGE	0.2718	0.3412	0.2013	0.1028	0.4972	0.3210	0.1132	0.4532	0.2647
NGCF	<b>0.2789</b>	<b>0.3463</b>	<b>0.2058</b>	<b>0.1065</b>	<b>0.5008</b>	<b>0.3226</b>	0.1156	0.4626	0.2732
MMGCN	<b>0.3057*</b>	<b>0.3996*</b>	<b>0.2298*</b>	<b>0.1164*</b>	<b>0.5520*</b>	<b>0.3423*</b>	<b>0.1250*</b>	<b>0.5186*</b>	<b>0.3085*</b>
%Improv.	<b>9.61%</b>	<b>15.59%</b>	<b>11.66%</b>	<b>9.03%</b>	<b>10.23%</b>	<b>6.11%</b>	<b>6.66%</b>	<b>9.78%</b>	<b>8.17%</b>



**Figure 3: Performance in terms of Recall@K w.r.t. different modalities on the three datasets.**

the concatenation of multi-modal features as content information to predict the interactions between users and micro-videos.

- **ACF** [8]. This is the first framework that is designed to tackle the implicit feedback in multimedia recommendation. It introduces two attention modules to address the item-level and component-level implicit feedbacks. To explore the modal-specific user preference and micro-video characteristic, we treat each modality as one component of the micro-video, which is consistent with the idea of standard ACF.
- **GraphSAGE** [14]. Such model is a general inductive framework that leverages node feature information to update node representations for the previously unseen data. In particular, it considers the structure information as well as the distribution of node features in the neighborhood. For a fair comparison, we integrate multi-modal features as the node features to learn the representation of each node.
- **NGCF** [41]. This method represent a novel recommendation framework to integrate the user-item interactions into the embedding process. By exploiting the higher-order connectivity from user-item interactions, the modal encodes the collaborative filtering signal into the representation. For a fair comparison, we regard the multi-modal features of micro-video as side information and feed them into the framework to predict the interactions between the users and items.

**Evaluation Protocols and Parameter Settings.** We randomly split the dataset into training, validation, and testing sets with 8:1:1 ratio, and create the training triples based on random negative sampling. For testing set, we pair each observed user-item pair with 1000 unobserved micro-videos that the user has not interacted before. We use the widely-used protocols [8, 19]: Precision@K, Recall@K, and NDCG@K to evaluate the performance of top-K

recommendation. Here we set  $K = 10$  and report the average scores in testing set. To train our proposed model, we randomly initialize model parameters with a Gaussian distribution and use the LeakyReLU as the activation function, and optimizing the model with stochastic gradient descent (SGD). We search the batch size in  $\{128, 256, 512\}$ , the latent feature dimension in  $\{32, 64, 128\}$ , the learning rate in  $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$  and the regularizer in  $\{0, 0.00001, 0.0001, 0.001, 0.01, 0.1\}$ . As the findings are consistent across the dimensions of latent vectors, if not specified, we only show the result of 64, a relatively large number that returns good accuracy.

### 3.2 Performances Comparison (RQ1)

The comparative results are summarized in Table 2. From this table, we have the following observations:

- MMGCN substantially outperforms all other baselines in most cases, verifying the effectiveness of our model. In particular, MMGCN improves the strongest baselines *w.r.t.* Recall by 15.59%, 10.23%, and 9.78%, respectively. We attribute such significant improvements to the learning of modal-specific representations, so as to capture users’ preference effectively.
- Jointly analyzing the results *w.r.t.* Recall on different datasets, we find that MMGCN has an advantage over the baselines on the three datasets. It again demonstrates that MMGCN is able to exploit the structural information to enrich the user and micro-video representations, which might benefit from the modal-specific user preference and micro-videos features modeling.
- The GNN-based model outperforms the CF-based model on Kwai and Tiktok. The improvements are attributed to the graph convolution layers. Such operations not only capture the local structure information but learn the distribution of neighbors’

**Table 3: Performance of MMGCN with different aggregation and combination layers.**

Variant	Kwai			Tiktok			MovieLens		
	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG
$g_{co-id}$	0.2812	0.3689	0.2146	0.1056	0.5289	0.3143	0.1034	0.4632	0.2702
$g_{co}$	0.2927	0.3841	0.2188	0.1132	0.5482	0.3372	0.1209	0.5090	0.3001
$g_{ele-id}$	0.2840	0.3729	0.2172	0.1071	0.5312	0.3186	0.1064	0.4704	0.2743
$g_{ele}$	<b>0.3057</b>	<b>0.3996</b>	<b>0.2298</b>	<b>0.1164</b>	<b>0.5520</b>	<b>0.3423</b>	<b>0.1250</b>	<b>0.5186</b>	<b>0.3085</b>

**Table 4: Performance of MMGCN w.r.t. the number of layers.**

Layer	Kwai			Tiktok			MovieLens		
	Precision	Recall	NDCG	Precision	Recall	NDCG	Precision	Recall	NDCG
One	0.2814	0.3728	0.2123	0.1084	0.5371	0.3263	0.1174	0.5017	0.2950
Two	<b>0.3057</b>	<b>0.3996</b>	<b>0.2298</b>	<b>0.1164</b>	<b>0.5520</b>	<b>0.3423</b>	<b>0.1250</b>	<b>0.5186</b>	<b>0.3085</b>
Three	0.2983	0.3910	0.2216	0.1103	0.5431	0.3361	0.1181	0.5032	0.2957

features for each ego node, boosting the expressiveness of representations.

- Generally speaking, NGCF achieves better performance than other baselines over three datasets in most cases. It is reasonable since NGCF are easily generalized to leverage the content information to characterize the users and micro-videos.
- Unexpectedly, ACF performs poor on all datasets. The reason of this may be due to the modification that we did during the implementation of the ACF model, in which we replaced the component-level features modeling by the modal-specific information for a fair comparison.

### 3.3 Study of MMGCN (RQ2)

**3.3.1 Effect of Modalities.** To explore the effect of different modalities, we compare the results on different modalities over the three datasets, as shown in Figure 3. It shows the performance of top- $K$  recommendation lists where  $K$  ranges from 1 to 10. From Figure 3, we have the following observations:

- As expected, the method with multi-modal features outperforms those with single-modal features in MMGCN on three datasets. It demonstrates that representing users with multi-modal information achieves higher accuracy. It further demonstrates that user representations are closely related to the content of items. Moreover, it shows that our model could capture the user’s modal-specific preference from content information.
- The visual modality is the most effective one among three modalities. It makes sense since, when a user chooses what to play, one usually pays more attention to the visual information than other modality information.
- The acoustic modality provides more important information for recommendation, compared with the textual features. In particular, for Tiktok dataset, the acoustic information even has comparable expressiveness to the visual modality.
- Textual modality is the least descriptive for interaction prediction, particularly on Kwai and Tiktok datasets. This is reasonable since we find the texts are of low quality, that is, the descriptions are noisy, incomplete, and even irrelevant to the micro-video content on these two datasets. However, this modality offers important cues on MovieLens dataset. Because the textual description is the storyline of the video, which highly relates to the content, and

some users may play the video according to the storyline. This phenomenon consists with our argument that the user preference are closely related to the content information.

- As  $K$  increases, Recall@ $K$  of MMGCN is consistently higher than the variants. It shows that user preference representations based on each modality are closer to the real preferences of users, which contribute to the prediction of user-item interactions. Modeling with user preferences on variety modalities can lead to quality multi-modal personalized recommendation.

**3.3.2 Effect of Combination Layers.** In our model, we design a novel combination layer to integrate the local structure information with the node’s features, facilitating the multiple modal-specific representation fusion. Wherein, the combination function can be implemented with two different way (cf. Equations (6) and (7)). Here we compare these different implementations and evaluate the effectiveness of the proposed combination layer, in which  $g_{co-id}$  and  $g_{ele-id}$  represent two type of implements without id embedding, respectively. As illustrated in Table 3, we have the following findings:

- In terms the three metrics, the  $g_{ele}$  one achieves the best performance on the three datasets. This may due to that the combination layer retains the modal-specific features to represent the users and micro-videos. It demonstrate the effectiveness of our combination layers.
- Comparing these methods, we found that the methods with id embedding significantly outperforms the others. This demonstrates the effectiveness of our novel combination layers. Besides, we suggest that the shared id embedding connects the different modalities by propagating the shared information during the backpropagation.
- Comparing the two implementations, we observed that the element-wise one is better than the concatenate one. We conjecture that the concatenate one with fully connected layer is more difficult to train, especially on the spare datasets, like Kwai.

**3.3.3 Effect of Model Depth.** To evaluate the effectiveness of layers stack, we conduct experiments on the three different layers, as shown in Table 4. From the results, we observed that:

- In terms three metrics, the two-layer model achieves better results, which shows the increasing of layers does not lead to better results. This demonstrates that the discrimination of



Figure 4: Visualization of users' played micro-videos distribution in different modalities, where each color indicates a user.

node is decreasing as the layers increasing. We suggest that the increasing layers makes the neighbors of nodes more similar, and further makes node representations more similar.

- Compared the one-layer model with the two-layer model, the improvement of the results on Tiktok and Kwai are more obvious, while the results on MovieLens are not significantly improved. It demonstrates that integrating the local structure information can enhance the node representation.
- Compared the two-layer model with the three-layer model, the later model achieved worse results than the former one. This may be caused by overfitting due to the sparse data.
- Compared the single layer model with the three-layer model, we observed that the results of single layer model are slightly inferior to those of the three-layer model. We suppose the insufficient local structure information of the single layer model results in the low quality node representation. This again demonstrates the effectiveness of content information in the node representations.

### 3.4 Case Study (RQ3)

We conducted experiments to visualize our modal-specific representations. In particular, we randomly sampled 5 users and collected the micro-videos they have played. To verify our assumption that the user preferences on different modalities are inconsistent, we visualized these representations using t-Distributed Stochastic Neighbor Embedding (t-SNE) in 2-dimension, as illustrated in Figure 4.

The coordinate graphs from left to right represent visual and textual modality, respectively. The points in the graphs represent the videos that the users have played, and their colors mean different users. Because the acoustic modality is hard to be visualized, we only analyze the results on visual and textual modalities.

- In visual modality, the points of user1 are dispersive and some of them mix with the points of user2, user3, and user4. The points form two concentrated regions in textual modality, and they are far apart from each other. The distribution of the points means he/she has two distinct preferred themes in textual modality, and they are quite different. While he/she has no particular preference on visual modality. The points of user2 centralized in three regions in visual modality. While in textual modality, they are diffuse and mix with other points

of users. The distribution pattern of user2 shows his/her three preferred themes in visual modality. The points of user3 are obviously centralized distribution in the two modalities, which illustrates he/she has particular preference in each modality. The distribution of the points of user4 and user5 are scattered and mix with other points of users.

- It is still abstract to use the distribution of points to analyze. The multi-modal information of videos represented by each point is displayed on the graph for further explanation. Take the example of user1 and user2, the visual and textual modality of some of their preferred videos are displayed in Figure 4, which represented by videos posters and storylines. We observed that the videos played by user1 have no obvious features visually, because the posters he/she preferred cover different varieties. However, the storylines of these videos are just including two themes: war and romance. From user2 we observed that his/her preference on visual modality are clearly divided into animation and classicism, while he/she has no distinct preference on storylines. These phenomenons demonstrate our assumption that the users have different preference in different modalities.

## 4 RELATED WORK

In this section, we introduce some works that are related to our researches, including multi-modal personalized recommendation, multi-modal fusion and graph convolution network.

### 4.1 Multi-modal Personalized Recommendation

Due to the success of CF method in recommendation systems, early multi-modal recommendation algorithms mainly based on CF models [6, 18, 38–40]. CF-based models leverage users' feedbacks (e.g. implicit feedback and explicit feedback) to predict the interactions between users and items. Although these approaches work well for the items with sufficient feedbacks, they are less applicable to the ones with few feedbacks, which cause the low-quality recommendations. Therefore, these CF-based methods are limited by the sparsity of the data.

To remedy the disadvantage of CF-based model, researchers have developed hybrid approaches which incorporate the items' content information and the collaborative filtering effects for

recommendation system. For instance, Chen *et al.* [7] constructed a user-video-query tripartite graph and performed graph propagation to combine the content information and the feedback information between users and videos. Recently, Chen *et al.* [8] explored the fine-grained user preference on the item and introduced a novel attention mechanism to address the challenging item- and component-level feedback in multimedia recommendation. In this method, the user is characterized by both collaborative filtering effect and the attended items' content information. Although this method has learned the two levels of the user preference, it fails to model the user preferences on different modalities, which is the key in the multi-modal recommendation as mentioned in Section 1. To fill up the blank in modal-specific features representation, our model constructs the graph in each modality and represents the model-specific features using GCN techniques, which integrates the local structure information and content information distribution of neighborhood.

## 4.2 Multi-modal Representation

The multi-modal representation is a representation of data using the information in multiple modalities, which is one of the most important problem in multi-modal applications. However, few works focus on the multi-modal representation in the area of multi-modal personalized recommendations.

Existing multi-modal representations can be grouped into two categories: joint representations and coordinated representations. Joint representations usually combine the single-modal information and project into the same representation space. The simplest implementation of the joint representation is the concatenation of single-modal features. Recently, with its success in computer vision [2, 15, 23] and natural language processing [9, 31], neural networks are increasingly used in the multi-modal domain, especially the multi-modal representations [10, 11, 34–37, 43]. Using the neural networks, the function fusing the different modalities information into a joint representation can be learned. Besides, the probabilistic graphical models [4, 13] are another way to construct a joint representation for multi-modal information using the latent random variable. Although these methods learn a joint representation to model the multi-modal data, they are suited for situations when all of the modalities are present during inference, which is hardly guaranteed in social platforms.

Different from the joint representations, the coordinated ones learn separate representations for each modality but coordinate them with constraints. To represent the multi-modal information, From *et al.* [12] proposed a deep visual-semantic embedding model which projects the visual information and semantic information into a common space constrained by distance between the visual embedding and corresponding word embedding. Similarly, Wang *et al.* [33] constructed a coordinated space which enforces images with similar meanings be closer to each other. However, since the modal-specific information is the factor causing the difference in each modality signals, the model-specific features are inevitably discarded via those similar constrains.

Differently, in our model, we introduced a novel representation, which respectively models the common part and specific part of features, to resolve the abovementioned problem.

## 4.3 Graph Convolution Network

As mentioned above, our proposed model uses the GCNs techniques to represent the users and micro-videos, which is widespread in recommendation systems [21, 22, 26, 28]. Towards the video recommendation, Hamilton *et al.* [14] proposed a general inductive framework which leverages the content information to generate node representation for unseen data. Based on this method, Ying *et al.* [42] developed and deployed a large-scale deep recommendation engine at Pinterest for image recommendation. In this model, the graph convolutions and random walks are combined to generate the representations of nodes. Currently, Berg *et al.* [5] treated the recommender systems as the view of link prediction on graphs and proposed a graph auto-encoder framework based on message passing on the bipartite interaction graph. Moreover, the side information can be integrated into the node representation via a separate processing channel. However, as can be seen, these methods fail to capture the modal-specific representation for each node in the multi-modal recommendation, which is the major concern of our work.

## 5 CONCLUSION AND FUTURE WORK

In this paper, we explicitly model modal-specific user preferences to enhance micro-video recommendation. We devised a novel GCN-based framework, termed MMGCN, to leverage information interchange between users and micro-videos in multiple modalities, refine their modal-specific representations, and further model users' fine-grained preferences on micro-videos. Experimental results on three public micro-video datasets well validated our model. In addition, we visualized some samples to illustrate the modal-specific user preferences.

This work investigates how the information exchange in different modalities influences user preference, which is an initial attempt to encode modality-aware structural information into representation learning. It would be a promising solution to understand user behaviors and provide more accurate, diverse, and explainable recommendation. In future, we will extend our MMGCN in several directions. First, we would construct multi-modal knowledge graph to present objects and relations between them in micro-videos [30], and then involve it into MMGCN to model finer-grained content analysis, explore user interests in a reasonable manner, and offer an in-depth understanding of user intents. It can provide more accurate, diverse, explainable recommendation. Second, we would explore how social leaders influence the recommendation, that is, integrating social network with user-item graphs. We also would like to incorporate multimedia recommendation into dialogue systems towards more intelligent conversational recommendations.

**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. This research is also part of NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

## REFERENCES

- [1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of International Conference on Learning Representations*. 1–16.
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2017), 2481–2495.
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.
- [4] Tadas Baltrušaitis, Ntombikayise Banda, and Peter Robinson. 2013. Dimensional affect recognition using continuous conditional random fields. In *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. IEEE, 1–8.
- [5] 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* (2017).
- [6] 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.
- [7] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. 2012. Personalized video recommendation through tripartite graph propagation. In *Proceedings of ACM international conference on Multimedia*. ACM, 1133–1136.
- [8] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 335–344.
- [9] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. In *Proceedings of ACM Multimedia Conference on Multimedia Conference*. ACM, 117–125.
- [10] 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 (TOIS)* 37, 2 (2019), 16.
- [11] Zhiyong Cheng, Shen Jialie, and Steven CH Hoi. 2016. On effective personalized music retrieval by exploring online user behaviors. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 125–134.
- [12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Proceedings of International Conference on Neural Information Processing Systems*. 2121–2129.
- [13] Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. 2008. Dynamic modality weighting for multi-stream hmms in audio-visual speech recognition. In *Proceedings of International Conference on Multimodal Interfaces*. ACM, 237–240.
- [14] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of International Conference on Neural Information Processing Systems*. 1024–1034.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [17] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1–8.
- [18] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 355–364.
- [19] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 173–182.
- [20] 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.
- [21] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. 2018. What Dress Fits Me Best?: Fashion Recommendation on the Clothing Style for Personal Body Shape. In *Proceedings of ACM Multimedia Conference on Multimedia Conference*. ACM, 438–446.
- [22] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *Proceedings of International Conference on Learning Representations*, 1–14.
- [23] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. 2019. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3838–3847.
- [24] 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.
- [25] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.
- [26] Federico Monti, Michael Bronstein, and Xavier Bresson. 2017. Geometric matrix completion with recurrent multi-graph neural networks. In *Proceedings of International Conference on Neural Information Processing Systems*. 3697–3707.
- [27] 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.
- [28] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *Proceedings of International conference on machine learning*. 2014–2023.
- [29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.
- [30] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. In *ICMR*. 279–287.
- [31] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. 160–170.
- [32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, and Yoshua Bengio. 2017. Graph Attention Networks. In *Proceedings of International Conference on Learning Representations*. 1–12.
- [33] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.
- [34] 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.
- [35] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and Tat-Seng Chua. 2012. Movie2comics: Towards a lively video content presentation. *IEEE Transactions on Multimedia* 14, 3 (2012), 858–870.
- [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] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. 2018. Joint Global and Co-Attentive Representation Learning for Image-Sentence Retrieval. In *Proceedings of ACM Multimedia Conference on Multimedia Conference*. ACM, 1398–1406.
- [38] 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 & Data Mining*.
- [39] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 1543–1552.
- [40] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item silk road: Recommending items from information domains to social users. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*. 185–194.
- [41] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*. 165–174.
- [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 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 974–983.
- [43] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems. In *The World Wide Web Conference*. ACM, 2401–2412.