Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for Personalized Micro-video Recommendation

Desheng Cai
Hefei University of Technology
caidsml@gmail.com

Shengsheng Qian
NLPR, Institute of Automation, CAS
The University of Chinese Academy of Sciences
shengsheng.qian@nlpr.ia.ac.cn

Quan Fang
NLPR, Institute of Automation, CAS
The University of Chinese Academy of Sciences
qfang@nlpr.ia.ac.cn

Jun Hu
NLPR, Institute of Automation, CAS
hujunxianligong@gmail.com

Changsheng Xu
NLPR, Institute of Automation, CAS
The University of Chinese Academy of Sciences
Peng Cheng Lab, ShenZhen, China
csxu@nlpr.ia.ac.cn

ABSTRACT

Micro-video recommendation has attracted extensive research attention with the increasing popularity of micro-video sharing platforms. There exists a substantial amount of excellent efforts made to the micro-video recommendation task. Recently, homogeneous (or heterogeneous) GNN-based approaches utilize graph convolutional operators (or meta-path based similarity measures) to learn meaningful representations for users and micro-videos and show promising performance for the micro-video recommendation task. However, these methods may suffer from the following problems: (1) fail to aggregate information from distant or long-range nodes; (2) ignore the varying intensity of users' preferences for different items in micro-video recommendations; (3) neglect the similarities of multi-modal contents of micro-videos for recommendation tasks. In this paper, we propose a novel Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for personalized micro-video recommendation. Specifically, we design a collaborative representation learning module and a semantic representation learning module to fully exploit user-video interaction information and the similarities of micro-videos, respectively. Furthermore, we utilize an anti-bottleneck module to automatically learn the importance weights of short-range and long-range neighboring nodes to obtain more expressive representations of users and micro-videos. Finally, to consider the varying intensity of users' preferences for different micro-videos, we design and optimize an adaptive recommendation loss to train our model in an end-to-end manner. We evaluate our method on three real-world datasets and the results demonstrate that the proposed model outperforms the baselines.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548420

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

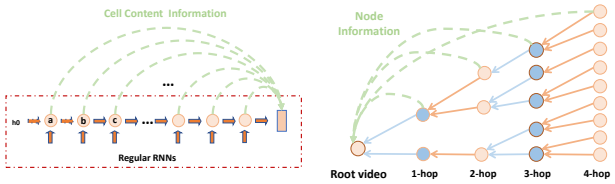
Multi-Modal, Bottleneck, Micro-video Recommendation

ACM Reference Format:
Desheng Cai, Shengsheng Qian, Quan Fang, Jun Hu, and Changsheng Xu. 2022. Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for Personalized Micro-video Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548420

1 INTRODUCTION

Personalized micro-video recommendation is an important yet challenging task. Most traditional approaches consider micro-video recommendation as a matching task [12, 50]. Basically, there are three major research lines: (1) Feature combination based methods, such as Factorization Machine (FM) [35] and deep Factorization Machine (deepFM) [8]; (2) ID-based recommendation approaches, such as collaborative filtering (CF) [42] and low-rank factorization [14]; (3) Multimedia recommendation methods, such as VBPR [10]. Although these algorithms show promising performance for recommendations, they rely heavily on multi-modal attribute information.

In recent years, graph representation learning method has become an emerging learning tool aiming to find meaningful representation for each node in the graph by effectively modeling the relationship between nodes. More precisely, Graph Neural Networks (GNNs) [9, 24, 44] have shown impressive performance in aggregating feature information of neighboring nodes, which may be leveraged to obtain more expressive representations of users and items for recommendation tasks (e.g. PinSAGE [51]). In addition, there are also many GNN-based approaches for personalized micro-video recommendation tasks. HGCL [1] leverages a heterogeneous graph encoder network and a graph contrastive learning network to capture the heterogeneity and global semantic information of user-video bipartite graph for micro-video recommendation. Therefore,



(a) RNNs (b) GNNs

Figure 1: Bottleneck problems of existing RNNs and GNNs.

it seems reasonable to improve the performance of micro-video recommendation by exploiting relationship information among users and micro-videos based upon existing GNN approaches.

However, these GNN-based methods still cannot effectively address the **following challenges** in the micro-video recommendation task: **(1)** Most existing graph-based recommendation methods heavily depend on graph convolutional networks (GNNs) based on user-item bipartite graphs to learn expressive representations of users and items for improving recommendation performance, while these GNN models are susceptible to the bottleneck problem when aggregating distant or long-range node information (e.g. HGCL [1]). Actually, the bottleneck problem of GNN models is analogous to the bottleneck problem of sequential RNN models. For example, in Fig 1(a), regular RNN models(e.g. traditional seq2seq models [5, 43]) have to contain the entire input sequence information into a fixed-length vector, which is likely to lead to a bottleneck problem for obtaining information from long-range nodes (e.g. a, b, c) in the receptive field of RNNs. These long-range nodes may fail to be compressed into target fixed-length node vectors. In RNNs, the receptive field of a node grows just linearly with the number of recursive operations. However in GNNs, the number of nodes in each node's receptive field grows exponentially as the number of GNN layers increases, as shown in Fig 1(b), which may make the bottleneck problems of GNNs asymptotically more harmful. For GNN-based recommendation tasks, GNNs may fail to aggregate messages flowing from long-range nodes, and learn only short-range node signals. **Therefore, how to break through the bottleneck problems of GNNs when aggregating distant or long-range node information for learning high-quality representations of users and micro-videos is a challenge in micro-video recommendation scenarios. (2)** Existing micro-video recommendation may fail to capture the varying intensity of users' preferences for different micro-videos by using the most used the weighted regression loss (WR) [19] or the Bayesian Personalized Ranking (BPR) [37] loss. For implicit data, the weighted regression loss aims to push similarity scores (normalized) of users with their all interactive items to 1, and push that of users with their all non-interactive items to 0. And the BPR loss optimizes recommendation models by modeling relative preferences. However, these losses may fail to model real distributions of similarity between users and interactive items (and non-interactive items). Actually, real distributions of similarity between users and items vary among different users, which, intuitively speaking, seems to the varying intensity of users' preferences for different items. We have discussed this in detail in section 4.5.2. **Therefore, how to optimize models by considering varying intensity of users' preferences for different micro-videos is a challenge for micro-video recommendations. (3)** Most existing GNN-based micro-video recommendation methods rely heavily on user-video historical interaction bipartite graphs for representation learning while ignoring the important role of their multi-modal content similarities in representation learning. As we all know, content similarities, especially micro-video multi-modal contents, are positively correlated with performances for recommendation tasks. However, these content relationships may not necessarily be captured in representation learning on user-item bipartite graphs due to issues such as missing interactive data or noise. **Therefore, how to learn representations by considering the similarities of multi-modal contents of micro-videos is a challenge for micro-video recommendation tasks.**

To address these challenges, in this paper, we propose a novel framework, named Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network (A^2BM^2GL), for personalized micro-video recommendation. **(1)** In order to fully exploit user-video interaction information and multi-modal content similarities of micro-videos, we design a collaborative representation learning module and a semantic representation learning module respectively. The collaborative representation learning module leverages GCNs on user-video interaction bipartite graphs to learn node representations. The semantic representation learning module can derive the k-nearest neighbor graph generated from multi-modal features as the feature graphs, which can effectively capture the underlying content similarities of micro-videos in the feature space. **(2)** To break through the bottleneck problems of GNNs when aggregating long-range node information, inspired by anti-bottleneck methods in RNNs as shown by green lines in Fig 1 (a), we design and utilize an anti-bottleneck module, which collects short-range neighboring nodes and long-range neighboring nodes from user-video interaction bipartite graph and video feature graphs to be a user-video interaction full-adjacent layer and several video feature full-adjacent layers respectively, as shown by green lines in Fig 1 (b). Then the attention mechanism is utilized to automatically learn the importance weights of short-range and long-range neighboring nodes jointly to obtain more expressive representations of users and micro-videos. **(3)** We fuse the learned representations from the collaborative and semantic representation learning module, and anti-bottleneck module to generate final representations of users and micro-videos. To consider the varying intensity of users' preferences for different items, we design an adaptive recommendation loss, which is an improved weighted regression and attempts to model real distributions of similarity scores between users and interactive micro-videos (and non-interactive micro-videos). Specifically, our loss tries to push similarity scores between each user and interactive micro-videos to an adaptive upper boundary which is learned to fit the mean value of the real distribution of similarity scores. Similarly, similarity scores between each user and non-interactive micro-videos are pushed to an adaptive lower boundary. Finally, we use a final unified loss to train our model in an end-to-end manner.

In summary, the contributions of this work are summarized: (1) We present a novel Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network (A^2BM^2GL) for personalized micro-video recommendations. Our model utilizes a collaborative representation learning module that applies GCNs on user-video interaction bipartite graphs to learn node representations and a semantic representation learning module that uses GCNs on video feature

graphs generated from multi-modal features to learn videos' representations; **(2)** An anti-bottleneck module is designed which utilizes the attention mechanism to automatically learn the importance weights of short-range and long-range neighboring nodes jointly in full-adjacent layers to obtain more expressive representations of users and micro-videos; **(3)** Our proposed model leverage an adaptive recommendation loss which is an improved weighted regression loss with adaptive upper boundary and lower boundary to consider the varying intensity of users' preferences for micro-videos; **(4)** We evaluate our method on the real-world datasets Kwai, Tiktok and MovieLens, and the results demonstrate that the proposed model outperforms baseline methods;

2 RELATED WORK

This work is closely related to micro-video recommendation, and graph neural networks, which are briefly reviewed.

Micro-video recommendation: Recommendation is the most effective tool to alleviate the information overload problem on video-sharing platforms. Existing approaches to deal with video recommendation can be roughly grouped into three categories: collaborative filtering [22, 25], content-based filtering [6, 32] and hybrid approaches [7, 21]. However, most of them are proposed to deal with traditional videos (e.g., videos on YouTube), while little work has been conducted for micro-videos. In recent years, with the increasing popularity of micro-video sharing platforms (e.g. Kwai and Tiktok), micro-video recommendation [3, 4, 20, 30] has attracted extensive research efforts to provide users with micro-videos in which they are interested. For example, Chen et al. [4] adopt a forward multi-head self-attention method with item-level and category-level attention module to model both behaviours and interests of users for micro-video recommendation. Recently, Liu et al. [27] use a stacked co-attention deep network to attend both user and video modalities and further model user and micro-video representations, which focus more on key features demonstrating users' hidden preferences. However, these models basically model users' preferences by utilizing users' historical interactions and the multi-modal content information of users and micro-videos.

Graph neural networks: The key idea behind graph neural networks (GNNs) is to aggregate feature information from node's local neighbor information by neural networks, which is widely used in many fields [16, 17, 33, 34]. GNNs can be utilized to represent the embeddings of users and items in structural graphs for recommendation tasks including micro-video recommendations [47–49, 51] For example, Wei et al. [49] propose a novel GCN-based framework, called MMGCN, leveraging information interchange between users and micro-videos to guide the representation learning in multiple modalities, and further model users' fine-grained preferences on micro-videos. Heterogeneous graphs (e.g. heterogeneous information networks) [38] can naturally model complex multiple types of objects and the rich relationships among them. There exit many heterogeneous graph based micro-video recommendation methods [1, 2, 15, 28, 40, 41, 53]. For example, Shi et al. [39] propose a novel heterogeneous network embedding based approach, called HERec, which can effectively integrate various kinds of embedding information based on different meta-paths in heterogeneous information network to enhance the personalized recommendation performance. Liu et al. [28] proposes a concept-aware denoising graph neural network (named Conde) for micro-video recommendation, which heavily relies on concept nodes, extracted from captions and comments of the videos, in a heterogeneous tripartite graph. HHFAN [2] is proposed to explore the highly complicated relationship from modality-aware heterogeneous information Graphs, and further generate high-quality noderepresentations based on a hierarchical feature aggregation network for recommendation. HGCL [1] leverages a heterogeneous graph encoder network and a graph contrastive learning network to obtain the better representations of users and micro-videos for better recommendation performances by capturing the heterogeneity and global semantic information of user-video bipartite graphs. However, these methods ignore bottleneck problems of GNN modules of recommendation models, and may not consider the varying intensity of users' preferences for different micro-videos effectively.

3 METHOD

3.1 Problem Statement

In this paper, we focus on micro-video recommendations, in which all micro-videos are denoted as $V = \{v_1, v_2, ..., v_{|V|}\}$ and all users are denoted as $U = \{u_1, u_2, ..., u_{|U|}\}$. Multi-modal attribute information (e.g., duration of micro-videos, micro-video visual content, micro-video audio content, micro-video textual description, the title of micro-video) of each micro-video is denoted as V^{attrs}: $V^{attrs} = \{v_1^a, v_2^a, ..., v_{|V_{attrs}|}^a\}$. We define a user-video interaction matrix as $I \in R^{|U|\times|V|}$, where the entry i_{uv} is defined from user's implicit feedback. Our purpose is to obtain corresponding high-quality representations of users and micro-videos given their multi-modal contents as well as interaction data, and calculate preference estimation scores to indicate users' preferences for recommendation.

3.2 Overall framework

In this work, we present a new framework, named Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network (A^2BM^2GL), for personalized micro-video recommendation, which, as shown in Figure 2, mainly consists of the following **four components:**

Collaborative Representation Learning Network: We design and utilize a collaborative representation learning module that applies GCNs on user-video historical interaction bipartite graphs to learn node representations. **Semantic Representation Learning Network:** To learn node representations of micro-videos by considering similarities of their multi-modal content features, we design a semantic representation learning module that can derive the k-nearest neighbor graph generated from multi-modal features as the feature graphs, and capture underlying content similarities of micro-videos in their feature space. **Anti-bottleneck Network:** We design an anti-bottleneck module that can automatically learn the importance weights of short-range and long-range neighboring nodes jointly in full-adjacent layers to obtain more expressive representations of users and micro-videos. **Adaptive Objective Functions and Optimization:** We design an adaptive recommendation loss which is an improved weighted regression loss with adaptive upper boundary and lower boundary to take the varying intensity of users' preferences for different micro-videos into consideration. In addition, we use our adaptive recommendation loss,

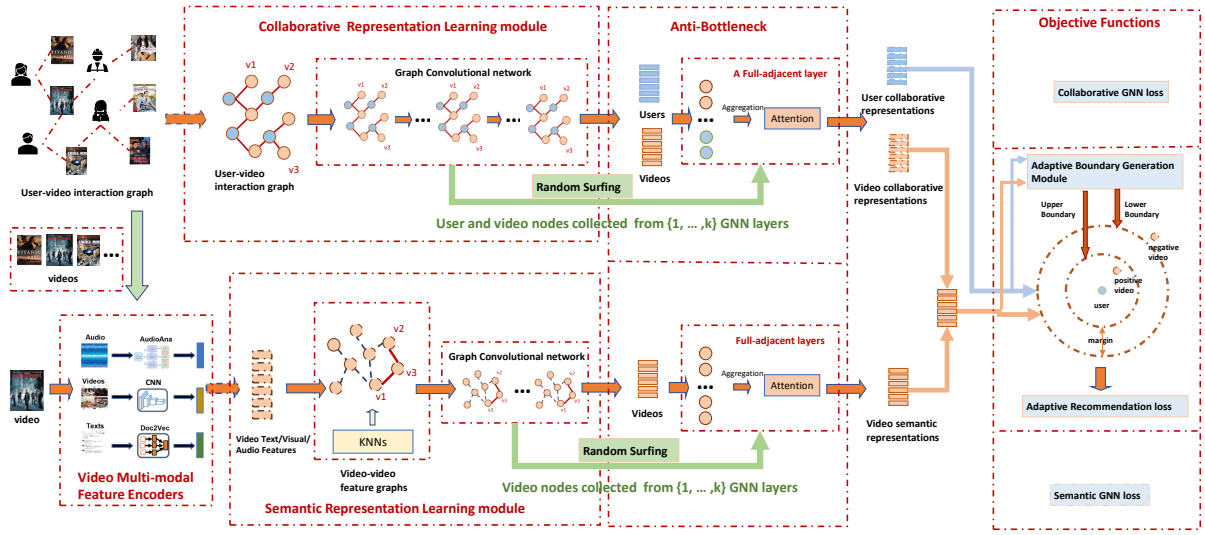


Figure 2: Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network (A^2BM^2GL).

collaborative GNN loss, and semantic GNN loss, to optimize our proposed framework jointly.

3.3 Collaborative Representation Learning Module

Firstly, we construct a user-video bipartite graph, defined as $G = \{U, V, A^{(uv)}\}$. In a bipartite graph G, U and V denote the set of users and micro-videos, respectively. $A^{(uv)}$ equals I and represents historical interaction relationships between users and micro-videos. Inspired by the LightGCN [11] model, we design and apply a collaborative representation learning network, which contains the simple weighted sum aggregator and abandons the use of feature transformation, to learn embeddings $\mathcal{E}$ of nodes of users and micro-videos. The graph convolution operation of users in a collaborative representation learning network is defined:

$$\mathcal{E}_u^{(k+1)} = \frac{1}{d_u + 1}\mathcal{E}_u^{(k)} + \sum_{v \in N_{(u)}} \frac{1}{\sqrt{|d_u + 1|}\sqrt{|d_v + 1|}}\mathcal{E}_v^{(k)} \quad (1)$$

where $N_{(u)}$ represents neighbor node sets of user u, d_u and d_v dnotes the original degree of the node, and $\frac{1}{\sqrt{|d_u+1|}\sqrt{|d_v+1|}}$ is the symmetric normalization term which follows the design of standard GCN [24]. To this end, we can obtain node representations of users and micro-videos based on collaborative representation learning network, which are $\mathcal{E}_u^{str}$ and $\mathcal{E}_v^{str}$ respectively. In order to learn more expressive representations of users and micro-videos, following the UltraGCN [31] model, we derive an extract loss:

$$\mathcal{L}_{str} = -\sum_{u \in U}\sum_{v \in N(u)} \beta_{u,v} log(\sigma(\mathcal{E}_u^T \mathcal{E}_v)), \quad \beta_{u,v} = \frac{1}{d_u}\sqrt{\frac{d_u + 1}{d_v + 1}} \quad (2)$$

where $N(u)$ is the set of interactive neighbor nodes of node u,

3.4 Semantic Representation Learning Network

Each micro-video is associated with some multi-modal attribute features, V_{attrs}. Video multi-modal attribute features can be pre-trained or initialized using different techniques $w.r.t.$ different types of attributes. We can utilize the doc2vec method [26] to pre-train attributes whose content type is text, employ CNN methods [29] to pre-train attributes whose content type is image and video, and utilize audioAna for embeddings of audio features. In order to capture the underlying structure of videos in video multi-modal feature space, we design a semantic representation learning network to construct multiple k-nearest neighbor (kNN) graphs based on video corresponding multi-modal feature matrices. Taking visual feature of videos as an example, $X^{(vis)} \in R^{|V| \times d}$ is visual feature matrix of videos. We aim to construct video visual feature graph $G^{(vis)} = (A^{(vis)}, X^{(vis)})$ using a similarity metric function, where $A^{(vis)}$ is the adjacency matrix of the KNN visual feature graph.

A good similarity metric function is supposed to be learnable and expressively powerful. Therefore, we design and utilize a weighted cosine similarity as our metric function as follows: $s_{ij} = cos(w \odot x_i, w \odot x_j)$, where x_i and x_j are visual feature vectors of video i and j, $\odot$ denotes the Hadamard product, and w is a learnable weight vector which has the same dimension as the input vectors x_i and x_j , and learns to highlight different dimensions of the vectors. To stabilize the learning process and increase the expressive power, we extend our similarity metric function to a multi-head version. Specifically, we use m weight vectors (each one representing one perspective) to compute m independent similarity matrices using the above similarity function and take their average as the final similarity as follows:

$$s_{ij}^p = cos(w_p \odot x_i, w_p \odot x_j), \quad s_{ij} = \frac{1}{m}\sum_{p=1}^{m} s_{ij}^p \quad (3)$$

Intuitively, s_{ij}^p computes the cosine similarity between the two input vectors x_i and x_j, for the p-th perspective, where each perspective considers one part of the semantics captured in the vectors.

After obtaining the similarity s_{ij} , and then we choose top similar node pairs for each node to set edges and finally get the adjacency matrix $A^{(vis)}$. To generate the more expressive video visual feature embedding matrix $H^{(vis)}$ in a graph $G^{(vis)}$, we use GCN encoder:

$$H^{(vis)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}^{(vis)}\hat{D}^{-\frac{1}{2}}X^{(vis)}W^{(vis)}) \tag{4}$$

where $\hat{A}^{(vis)} = A^{(vis)} + wI_n$ which is the adjacency matrix $A^{(vis)}$ of the graph $G^{(vis)}$ with added self-connections, $\hat{D}_i = \sum_j \hat{A}_{ij}^{(vis)}$ which stands for the degree matrix of $\hat{A}^{(vis)}$, $W^{(vis)} \in R^{f\times d}$ is a trainable weight matrix of video visual feature graph encoder, and σ is the ReLU nonlinearity. Unlike conventional GCNs, we control the weight of the self-connections by introducing a weight $w \in R$. Larger w indicates that the node itself plays a more important role in generating its embedding, which in turn diminishes the importance of its neighboring nodes. In addition, following the UltraGCN model, we also derive an extract loss $\mathcal{L}_{sem}$ for a semantic representation learning network similar to a collaborative representation learning module. Based on the semantic representation learning network, we can learn visual content representations $\mathcal{E}_v^{vis}$ of micro-videos. Similarly, we can obtain other modal content representations of micro-videos, such as $\mathcal{E}_v^{text}$, $\mathcal{E}_v^{audio}$ and so on.

3.5 Anti-bottleneck Network

Although we can obtain representations of users and micro-videos by graph encoder networks from collaborative representation learning networks and semantic representation learning networks, these graph encoder networks still suffer from bottleneck problems, which may lead to less ideal recommendation performances. Therefore, we design and utilize an anti-bottleneck module, which collects short-range neighboring nodes and long-range neighboring nodes from user-video interaction bipartite graph and video feature graphs to be a user-video interaction full-adjacent layer and several video feature full-adjacent layers respectively. To obtain more meaningful full-adjacent layers, we utilize the random surfing model by weighting the importance of contextual nodes based on their relative distance to the current node. Specifically, following the HGCL model, the key part of the random surfing model is the transition operation with a restart function: a probability α denotes that the transition operation will continue, and a probability $1-\alpha$ denotes that it will return to the original node and restart the transition operation. Taking the user-video bipartite graph as an example, this leads to the following recurrence relation:

$$A_k = \alpha A_{k-1}A + (1-\alpha)A_0, \quad A = \begin{pmatrix} 0 & A^{(uv)} \\ (A^{(uv)})^T & 0 \end{pmatrix} \tag{5}$$

where A is a direct user-video transition matrix, and A_k is the result after k transition operations, A_0 is an identity matrix. We define a matrix $M^{(uv)}$ to preserve all of transition results: $M^{(uv)} = \sum_k A_k$. Further, we take $M^{(uv)}$ as the unnormalized adjacency matrix in the user-video interaction full-adjacent layer. Similarly, we can obtain adjacency matrices in video feature full-adjacent layers, such as $M^{(vis)}$, $M^{(text)}$, and so on.

In order to aggregate node representations from full-adjacent layers M into combined representations of nodes by considering the different impacts of these representations on current nodes, we employ an attention technique based on representations (e.g. $\mathcal{E}_u^{str}, \mathcal{E}_v^{vis}, \mathcal{E}_v^{text}$) learned from the collaborative representation learning network and the semantic representation learning network. Take the micro-video visual content full-adjacent layer $M^{(vis)}$ as an example, we compute attention values for node v as follows:

$$\alpha^{v,i} = \frac{exp(LeakyReLU(w^T[\mathcal{E}_v^{(vis)} \bigoplus \mathcal{E}_{v,i}^{(vis)}))}{\sum exp(LeakyReLU(w^T[\mathcal{E}_v^{(vis)} \bigoplus \mathcal{E}_v^{(vis)}]))} \tag{6}$$

where $\alpha^{v,*}$ indicates the importance of different representations and w^T is the attention parameter. $\mathcal{E}_v^{(anti\_vis)}$ is the combined representation of node v: $\mathcal{E}_v^{(anti\_vis)} = \sum_{N_v} \alpha^{v,i}\mathcal{E}_{v,i}^{(vis)}$. N_v is the set of neighbor nodes in $M^{(vis)}$ of node v. Similarly, we can obtain other combined representations of nodes from our anti-bottleneck module, such as $\mathcal{E}_u^{(anti\_str)}$, $\mathcal{E}_v^{(anti\_str)}$, $\mathcal{E}_v^{(anti\_text)}$ and so on.

3.6 Adaptive Loss and Optimization

3.6.1 ***Representation Fusion Module.*** From the previous part, representations of users and micro-videos from the collaborative representation learning module, the semantic representation learning module, and anti-bottleneck module are generated. In order to fuse these representations into the final representation of users and micro-videos, we just apply sum function for users: $\mathcal{E}_u = \sum_i \mathcal{E}_u^i$ Similarly, we obtain final representations of micro-videos, $\mathcal{E}_v$.

3.6.2 ***Adaptive Micro-video Recommendation Loss.*** For each user, the main goal of the framework is to predict a list of micro-videos that a user may prefer. Therefore, given a user u and a candidate micro-video v, we compute a predicted preference probability $\hat{y}$ based upon their learned representations. Formally, we calculate the preference probability: $\hat{y}_{uv} = \sigma(\mathcal{E}_u{}^T\mathcal{E}_v)$, where $\mathcal{E}_u$ and $\mathcal{E}_v$ are learned representations of user u and candidate micro-video v respectively. To effectively take the varying intensity of users' preferences for different items into consideration, we leverage an adaptive recommendation loss which is an improved weighted regression loss with adaptive upper boundaries and lower boundaries. In particular, we utilize a non-sampling strategy for our loss. Firstly, we formulate the weighted regression (WR) recommendation loss function, which treats all missing interactions as negative instances and weighting them uniformly, $\mathcal{L}_{WR}$ as follows:

$$\mathcal{L}_{WR} = \sum_{u\in U}(\sum_{v\in N(u)^+} c_{uv}(1-\hat{y}_{uv})^2 + \sum_{v'\in N(u)^-} c_{uv'}(\hat{y}_{uv'}-0)^2) \tag{7}$$

where c_{uv} denotes the weight of entry (u,v), $N(u)^+$ is the set of interactive neighbor nodes of node u, and $N(u)^-$ is the set of non-interactive nodes of node u, However, the weighted regression loss aims to push predicted probabilities of users with their all interactive items to 1, and push that of users with their all non-interactive items to 0. It seems that the weighted regression loss cannot effectively reflect the real distribution of predicted probability. Therefore, we design an adaptive recommendation loss for modeling real distributions of similarity scores as follows:

$$\mathcal{L}_{rec} = \sum_{u\in U}(\sum_{v\in N(u)^+}(B_u^{up} - \mathcal{E}_u{}^T\mathcal{E}_v)^2 + \sum_{v\in N(u)^-}(\mathcal{E}_u{}^T\mathcal{E}_v - B_u^{low})^2) \tag{8}$$

Our loss tries to push similarity scores between user u and all interactive micro-videos to an adaptive upper boundary B_u^{up} which is learned to fit the mean value of the real distribution of similarity scores. Similarly, B_u^{low} is learned to fit the mean value of the real distribution of similarity scores between user u and all non-interactive micro-videos. For each user u, we calculate an adaptive

Table 1: Statistics of Datasets.

Dataset	User	micro-video	Interaction	Sparsity
Kwai	10000	34325	4,221,975	98.75%
Tiktok	15,727	153,005	215,942	99.99%
MovieLens	943	1682	10,000	93.7%

upper boundary B_u^{up} and an adaptive lower boundary B_u^{low} by the adaptive boundary generation module as follows:

$$B_u^{up} = MLP\{\mathcal{E}_u, \sum_{v \in N(u)^+} \mathcal{E}_v\}, \quad B_u^{low} = MLP\{\mathcal{E}_u, \sum_{v \in N(u)^-} \mathcal{E}_v\} \quad (9)$$

To constrain B_u^{up} to be greater than B_u^{low} and have a certain size of margin m, we use an additional loss: $\mathcal{L}_{margin} = -max\{m - (B_u^{up} - B_u^{low}), 0\}$. Combined with previous losses, the final loss function $\mathcal{L}_{final}$ becomes as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{rec} + \delta\mathcal{L}_{margin} + \beta\mathcal{L}_{str} + \gamma\mathcal{L}_{sem} + \frac{1}{2}\lambda||\theta||_2^2 \quad (10)$$

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

4 EXPERIMENTS

4.1 Datasets

We conduct experiments on three real-world datasets: Kwai dataset ,Tiktok dataset and MovieLens (MLs) dataset. We summarize the statistics of datasets in Table 1. **Kwai dataset**[1]**:** It is a real-world dataset collected from the Kwai platform. Micro-videos are associated with multi-modal features. **Tiktok dataset**[2]**:** It is released by Tiktok, a micro-video sharing platform. Note that attributes of micro-videos are multi-modal features. **MovieLens (MLs) dataset**[3]**:** We use a 1M version of the dataset and we collect multi-modal information according to the movie names provided in the dataset and further construct multi-modal features for experiments.

4.2 Baselines

We use the following micro-video recommendation algorithms as baselines to establish a fair comparison and illustrate the effectiveness of our proposed model. Traditional methods: **FM** [36] and **DeepFM** [8]; Deep neural networks: **NeuMF** [13]; Homogeneous GNN based networks: **PinSAGE** [51] and **GraphSAGE** [9]; Heterogeneous GNN based networks: **MMGCN** [49], **HeRec** [39], **SemRec** [41], **HAN** [46], and **HetGNN** [52]; Homogeneous structure learning based networks: **IDGL** [45]; The state-of-the-art method: **HGCL** [1] utilizes a graph contrastive learning network on each node-type specific homogeneous graph, which enables the learned embeddings of users and micro-videos to better capture the heterogeneity of bipartite interaction graph and global information of each homogeneous graph for micro-video recommendation.

4.3 Evaluation Metrics and Parameter Settings

For each dataset, we randomly hold 80% of micro-videos associated with each user to construct training datasets, and use remaining micro-videos to construct testing datasets. We employ cross-validation with grid-search to tune all hyper-parameters for searching the best model. We evaluate our method and baselines on micro-video recommendation by using four popular Top-K recommendation metrics [49], Precision at top K (P@K), Recall at top K (R@K), Normalized Discounted Cumulative Gain at top K (NDCG@K) and AUC. Here we set K = 10 and report the average results in the testing datasets. To train our proposed model, we use a Gaussian distribution to randomly initialize the model parameters, and optimize our model through mini-batch Adaptive Moment Estimation (Adam) [23]. We search the dimensions of latent vectors d in {100, 150, 200, 250, 300, 350}, the learning rate in {1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2} and the regularizer in {0, 1e-5, 1e-4, 1e-3, 1e-2,, 1e-6}. We set δ = 0.5, β = 0.6 and γ = 0.1.

4.4 Performances Comparison

The comparative results are summarized in Table 2. From the results, we have the following observations: (1) Homogeneous GNN based networks (GraphSAGE and PinSAGE) outperform traditional methods (FM and DeepFM) and deep neural networks (NeuMF) on three datasets. This shows that graph convolution operations can capture graph structure information and neighbors' multi-modal features for each node, which can improve the quality of representations for micro-video recommendation. Heterogeneous information graphs based baselines (MMGCN, SemRec, HeRec and HetGNN), which leverage message propagation mechanism to learn node embeddings in heterogeneous graphs, achieve better results than other baselines over three datasets in most cases. This indicates that the heterogeneity of data, especially for graphs, is also important for micro-video recommendation. (2) Compared with heterogeneous information graphs based baselines, HGCL achieves better performances. This indicates that not only considering the heterogeneity of user-video bipartite graphs but also the non-local (global) semantic information user-video bipartite graphs are very useful for improving the performance of micro-video recommendation tasks. IDGL learns representations of micro-videos based on their multi-modal feature graphs, result of which are similar to that of heterogeneous information graphs based baselines. However, the bottleneck problem of GNN is not considered. (3) Our proposed model consistently beats all the baselines in all cases, verifying the effectiveness of our model for micro-video recommendation. Compared with HGCL, the state-of-the-art micro-video recommendation method, our method achieves meaningful improvements in terms of all metrics on three datasets as shown in Table 2, and has the following advantages: Our proposed model utilizes a collaborative representation learning module, and a semantic representation learning module to learn representations which can learn useful representations on user-video bipartite graphs and video feature graphs. In addition, an anti-bottleneck module can effectively break through bottleneck problems of previous GNN modules by utilizing the attention mechanism to automatically learn the importance weights of short-range and long-range neighboring nodes jointly constructed full-adjacent layers. Finally, our proposed model leverage an adaptive recommendation loss, which can effectively capture the varying intensity of users' preferences for different micro-videos for improving recommendation performances.

[1]https://www.kwai.com/
[2]http://ai-lab-challenge.bytedance.com/tce/vc/
[3]http://grouplens.org/datasets/movielens/1m/

Table 2: Performance comparisons between our model and the baselines on three datasets. (K=10, training_rate=0.8)

Datasets	Metrics	FM	DeepFM	NeuMF	GraphSage	PinSage	MMGCN	SemRec	HeRec	HAN	HetGNN	IDGL	HGCL	A^2BM^2GL
Kwai	Pre	0.0447	0.0435	0.0411	0.0844	0.0822	0.0795	0.0995	0.1043	0.1012	0.1051	0.1006	0.1131	**0.1291**
	Rec	0.0143	0.0062	0.0111	0.0151	0.0127	0.0113	0.0131	0.0165	0.0141	0.0145	0.0155	0.0253	**0.0281**
	NDCG	0.0615	0.0664	0.0417	0.1025	0.1011	0.0931	0.1040	0.1081	0.1121	0.1139	0.1001	0.1213	**0.1398**
	AUC	0.6681	0.6701	0.6555	0.6985	0.7213	0.7130	0.7011	0.7077	0.7220	0.7212	0.7003	0.7401	**0.7610**
Tiktok	Pre	0.2688	0.2531	0.1929	0.3433	0.3328	0.3423	0.3327	0.3631	0.3655	0.3813	0.3645	0.3921	**0.4117**
	Rec	0.2243	0.2091	0.1533	0.2637	0.2515	0.2015	0.2695	0.2723	0.3029	0.3143	0.3012	0.3321	**0.3499**
	NDCG	0.2721	0.3029	0.2912	0.3585	0.3713	0.3015	0.3815	0.3778	0.3899	0.3835	0.3763	0.4024	**0.4219**
	AUC	0.6667	0.6431	0.6534	0.7050	0.7095	0.6989	0.7222	0.7187	0.7233	0.7289	0.7113	0.7435	**0.7720**
MLs	Pre	0.1568	0.1559	0.2020	0.2988	0.3039	0.2545	0.3009	0.3225	0.3129	0.3321	0.3099	0.4016	**0.4295**
	Rec	0.1514	0.1888	0.1733	0.2297	0.1988	0.2018	0.2410	0.2392	0.2141	0.2434	0.2212	0.2651	**0.2833**
	NDCG	0.3231	0.3005	0.3511	0.3799	0.3525	0.3179	0.4099	0.4411	0.4508	0.4527	0.4121	0.4799	**0.5006**
	AUC	0.6412	0.6549	0.7016	0.7211	0.7235	0.6937	0.7331	0.7433	0.7341	0.7533	0.7241	0.7621	**0.7834**

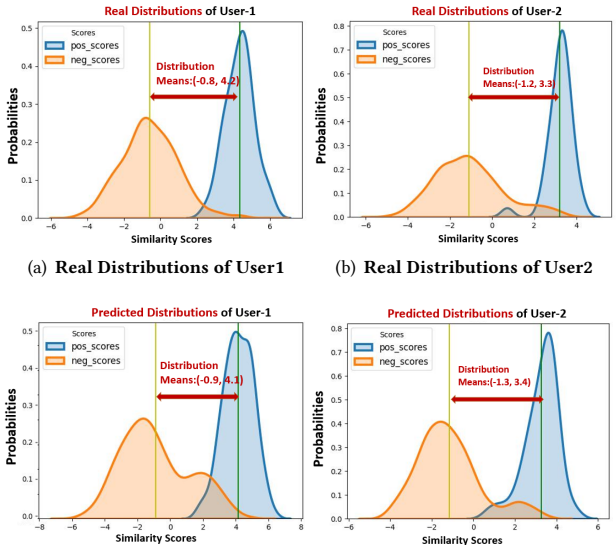
Table 3: Performance comparisons between model variants on three datasets. (K=10)

Datasets	Metrics	model¬an	model¬al	model¬str	model
Kwai	NDCG@10	0.1114	0.1225	0.1073	**0.1398**
	AUC	0.7328	0.7499	0.7322	**0.7610**
Tiktok	NDCG@10	0.3959	0.4018	0.4001	**0.4219**
	AUC	0.7375	0.7569	0.7241	**0.7720**
MLs	NDCG@10	0.4602	0.4801	0.4541	**0.5006**
	AUC	0.7364	0.7622	0.7321	**0.7834**

4.5 Detail Analysis

4.5.1 ***Ablation study.*** Since the proposed model contains multiple key components, we compare the following variants of model to demonstrate the effectiveness of the proposed model: (1) **model**¬*an*: It is a variant of our model, which removes the anti-bottleneck module. (2) **model**¬*al*: It is a variant of our model, which removes our designed adaptive recommendation loss function, and use the BPR loss instead. (3) **model**¬*str*: It is a variant of our model, which removes the semantic representation learning module. The ablation study results are reported in Table 3. From the results, we can conclude that: (1) The anti-bottleneck module is designed to break through bottleneck problems existed in collaborative and semantic representation learning modules. Our model has better performance than model¬*an* in all cases, demonstrating that the anti-bottleneck module can effectively improve the representation learning ability of collaborative and semantic representation learning modules. (2) Our designed adaptive recommendation loss aims to model real distributions of similarity of users with their positive micro-videos (and their negative micro-videos). Our model outperforms the model¬*al*, showing that our recommendation loss can learn the real similarity distributions between different samples more effectively. (3) The collaborative representation learning module focuses on capturing similarities between contents of micro-videos in each content modality(e.g. text, visual) for micro-video representation learning. Our model is superior to that of model¬*str*, indicating that our model can effectively obtain high-quality representations of micro-videos based on multi-model content features.

4.5.2 ***Distribution visualization.*** In order to better demonstrate the superiority of our proposed loss, we visualize the distribution



(a) **Real Distributions of User1** (b) **Real Distributions of User2**

(c) **Predicted Distributions of User1** (d) **Predicted Distributions of User2**

Figure 3: Real and predicted score distributions in terms of positive items and negative items for different users

results of similarity between different users and their interactive micro-videos(and non-interactive micro-videos). As shown in Fig3, under the ideal condition where we train the model with real data distribution (training set combined with test set) and output test results, Fig3(a) and Fig3(a) visualize real distributions of similarity scores between users and all their interactive micro-videos(pos_scores) and non-interactive micro-videos(neg_scores) in terms of user-1 and user-2, respectively. In Fig3(a), real distributions of user-1 and interactive micro-videos(pos_scores) and non-interactive micro-videos(neg_scores) presents normal distributions, with mean values of -0.8 and 4.2. In Fig3(b), real distribution of user-2 is with mean values of -1.2 and 3.3, respectively. Existing losses fail to model real distributions of similarity between users and interactive items (and non-interactive items). Our loss tries to push similarity scores between each user and interactive micro-videos to an adaptive upper boundary which is learned to fit the mean value of the real distribution of similarity scores. Under normal circumstances where we only train the model with the training

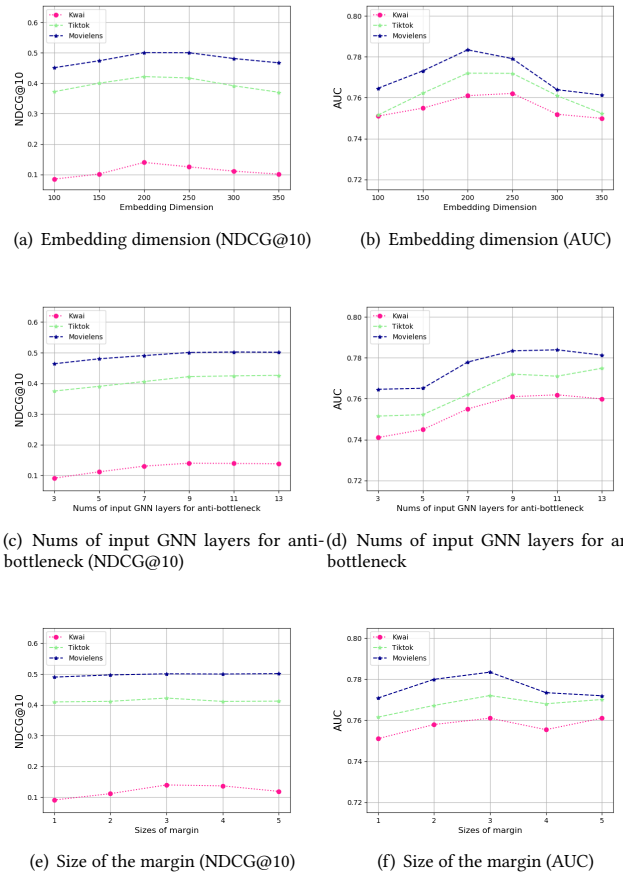
Table 4: Performance comparisons for different losses on three datasets. (K=10)

Datasets	Metrics	WR	BPR	InfoBPR	Our loss
Kwai	NDCG@10	0.1173	0.1225	0.1274	**0.1398**
	AUC	0.7411	0.7499	0.7502	**0.7610**
Tiktok	NDCG@10	0.3771	0.4018	0.4137	**0.4219**
	AUC	0.7319	0.7569	0.7614	**0.7720**
MLs	NDCG@10	0.4598	0.4801	0.4951	**0.5006**
	AUC	0.7522	0.7622	0.7736	**0.7834**

set and output the test results. Fig3(c) and Fig3(d) respectively show prediction distributions of similarity scores for user-1 and user-2, showing that the prediction distributions of our model are very similar to those shown by Fig3(a) and Fig3(b). The mean values of the normal distribution in Fig3(c) are -0.9 and 4.1 respectively (the ideal results are -0.8 and 4.2 in Fig3(a). The mean values of the normal distribution in Fig3(d) are -1.3 and 3.4 respectively (the ideal results are -1.2 and 3.3 in Fig3(b). We can conclude that our model has achieved good results.

4.5.3 ***Impact of different losses.*** In order to further prove the effectiveness of our proposed loss, we change our model into other commonly used recommended objective functions, which include the weighted regression (WR) loss, the BPR loss and the infoBPR [18] loss which is an improved BPR loss, and present the experimental results in Table 4. According to the experimental results, the performance of our loss is the best showing that our loss can more reasonably model the similarity distribution between users and interactive micro-videos (and non-interactive micro-videos).

4.5.4 ***Hyper-parameters sensitivity.*** We conduct experiments to analyze the impacts of three key parameters of the proposed model including the embedding dimension d of users and micro-videos, nums of input GNN layers for the anti-bottleneck module and the size of the margin m. The results are shown in Figure 4. From the results, we have the following observations: (1) When the embedding dimension d of each node varies from 100 to 200, NDCG@10 and AUC are basically increasing. However, when the dimension d gets larger, the performance of our model declines to a certain extent, which may be due to the overfitting of our model. (2) When the number k of GNN layers of anti-bottleneck module input nodes varies from 3 to 9, results of AUC and NDCG@10 rise steadily to the optimum. In addition, when the number of k grows larger, the performance is basically stable. This may prove to some extent that our model can break through the limitation of the bottleneck of GNN modules. (3) As size of the margin gradually becomes larger, the performance of our model is improving. Then, when the margin is larger, our model's performance deteriorates. The reason is that large or small margins may not conform to the real data distribution. Note that the best margins may vary for different datasets.



(a) Embedding dimension (NDCG@10) (b) Embedding dimension (AUC)

(c) Nums of input GNN layers for anti-bottleneck (NDCG@10) (d) Nums of input GNN layers for anti-bottleneck

(e) Size of the margin (NDCG@10) (f) Size of the margin (AUC)

Figure 4: Impacts of different parameters.

5 CONCLUSIONS

In this paper, we investigate the problem of personalized micro-video recommendation. We argue that most GNN-based recommendation approaches ignore bottleneck problems that existed in GNN modules of recommendation models, and may not consider the varying intensity of users' preferences for different micro-videos effectively. In this paper, we propose a novel framework, named Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network (A^2BM^2GL), for personalized micro-video recommendation. Specifically, our proposed model utilizes a collaborative representation learning module to learn node representations and a semantic representation learning module on video feature graphs to learn videos' representations. In addition, an anti-bottleneck module utilizes the attention mechanism to automatically learn the importance weights of short-range and long-range neighboring nodes jointly to obtain more expressive representations. Finally, our proposed model leverages an adaptive recommendation loss with adaptive upper boundary and lower boundary to consider the varying intensity of users' preferences for items. Experimental results on three micro-video datasets show that our algorithm outperforms existing methods in the micro-video recommendation task.

6 ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (No.62036012, 61936005, 61832002, 61721004, 62072456, 62106262, 61872199), the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJSSWJSC039, and the Open Research Projects of Zhejiang Lab (NO.2021KE0AB05).

REFERENCES

[1] Desheng Cai, Shengsheng Qian, Quan Fang, Jun Hu, Wenkui Ding, and Changsheng Xu. 2022. Heterogeneous Graph Contrastive Learning Network for Personalized Micro-video Recommendation. *IEEE Transactions on Multimedia* (2022).

[2] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-Video Recommendation. *IEEE Transactions on Multimedia* (2022).

[3] 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 *SIGIR 2017*. 335–344.

[4] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 1146–1153.

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1724–1734.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*. 191–198.

[7] Andrea Ferracani, Daniele Pezzatini, Marco Bertini, and Alberto Del Bimbo. 2016. Item-Based Video Recommendation: An Hybrid Approach considering Human Factors. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*. 351–354.

[8] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 1725–1731.

[9] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 1024–1034.

[10] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 144–150.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. *LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.* 639–648.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW 2017*. 173–182.

[13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 173–182.

[14] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. 549–558.

[15] Binbin Hu, Chuan Shi, Wayne Xin Zhao, and Philip S. Yu. 2018. Leveraging Meta-path based Context for Top- N Recommendation with A Neural Co-Attention Model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 1531–1540.

[16] Jun Hu, Shengsheng Qian, Quan Fang, Youze Wang, Quan Zhao, Huaiwen Zhang, and Changsheng Xu. 2021. Efficient graph deep learning in tensorflow with tf_geometric. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3775–3778.

[17] Jun Hu, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Hierarchical Graph Semantic Pooling Network for Multi-modal Community Question Answer Matching. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1157–1165.

[18] Jun Hu, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. MGDCF: Distance Learning via Markov Graph Diffusion for Neural Collaborative Filtering.

[19] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*.

[20] Lei Huang and Bin Luo. 2017. Personalized Micro-Video Recommendation via Hierarchical User Interest Modeling. In *Advances in Multimedia Information Processing - PCM 2017 - 18th Pacific-Rim Conference on Multimedia, Harbin, China, September 28-29, 2017, Revised Selected Papers, Part I (Lecture Notes in Computer Science, Vol. 10735)*. 564–574.

[21] Qinghua Huang, Bisheng Chen, Jingdong Wang, and Tao Mei. 2014. Personalized Video Recommendation through Graph Propagation. *TOMCCAP* 10, 4 (2014), 32:1–32:17.

[22] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. 2016. Real-time Video Recommendation Exploration. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*. 35–46.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[24] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[25] Noam Koenigstein and Ulrich Paquet. 2013. Xbox movies recommendations: variational bayes matrix factorization with embedded feature selection. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*. 129–136.

[26] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings, Vol. 32)*. 1188–1196.

[27] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-Video Co-Attention Network for Personalized Micro-video Recommendation. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. 3020–3026.

[28] Yiyu Liu, Qian Liu, Yu Tian, Changping Wang, Yanan Niu, Yang Song, and Chenliang Li. 2021. Concept-Aware Denoising Graph Neural Network for Micro-Video Recommendation. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. 1099–1108.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 3431–3440.

[30] Jingwei Ma, Guang Li, Mingyang Zhong, Xin Zhao, Lei Zhu, and Xue Li. 2018. LGA: latent genre aware micro-video recommendation on social media. *Multimedia Tools Appl.* 77, 3 (2018), 2991–3008.

[31] K. Mao, J. Zhu, X. Xiao, B. Lu, Z. Wang, and X. He. 2021. UltraGCN: Ultra Simplification of Graph Convolutional Networks for Recommendation. (2021).

[32] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual Video Recommendation by Multimodal Relevance and User Feedback. *ACM Trans. Inf. Syst.* 29, 2 (2011), 10:1–10:24.

[33] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[34] Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2440–2448.

[35] Steffen Rendle. 2010. Factorization Machines. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*. 995–1000.

[36] Steffen Rendle. 2012. Factorization Machines with libFM. *ACM TIST* 3, 3 (2012), 57:1–57:22.

[37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. 452–461.

[38] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and Philip S. Yu. 2019. Heterogeneous Information Network Embedding for Recommendation. *IEEE Trans. Knowl. Data Eng.* 31, 2 (2019), 357–370.

[39] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and Philip S. Yu. 2019. Heterogeneous Information Network Embedding for Recommendation. *IEEE Trans. Knowl. Data Eng.* 31, 2 (2019), 357–370.

[40] Chuan Shi, Jian Liu, Fuzhen Zhuang, Philip S. Yu, and Bin Wu. 2016. Integrating heterogeneous information via flexible regularization framework for recommendation. *Knowl. Inf. Syst.* 49, 3 (2016), 835–859.

[41] Chuan Shi, Zhiqiang Zhang, Yugang Ji, Weipeng Wang, Philip S. Yu, and Zhiping Shi. 2019. SemRec: a personalized semantic recommendation method based on weighted heterogeneous information networks. *World Wide Web* 22, 1 (2019), 153–184.

[42] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Adv. Artificial Intellegence* 2009 (2009), 421425:1–421425:19.

[43] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 3104–3112.

[44] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

[45] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[46] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. 2022–2032.

[47] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[48] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.

[49] 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 International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. 1437–1445.

[50] Jun Xu, Xiangnan He, and Hang Li. 2018. Deep Learning for Matching in Search and Recommendation. In *SIGIR 2018*. 1365–1368.

[51] 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 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 974–983.

[52] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. 2019. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. 793–803.

[53] Jing Zheng, Jian Liu, Chuan Shi, Fuzhen Zhuang, Jingzhi Li, and Bin Wu. 2017. Recommendation in heterogeneous information network via dual similarity regularization. *Int. J. Data Sci. Anal.* 3, 1 (2017), 35–48.