Pattern Recognition 136 (2023) 109211

Contents lists available at ScienceDirect

Pattern Recognition

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

Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks

Fei Wu[a], Shuaishuai Li[a], Guangwei Gao[c,*], Yimu Ji[b,d], Xiao-Yuan Jing[e,f,*], Zhiguo Wan[g]

[a] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, Hangzhou, China
[c] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China
[d] College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China
[e] School of Computer, Wuhan University, Wuhan, China
[f] Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis and School of Computer, Guangdong University of Petrochemical Technology, Maoming, China
[g] Basic Theory Research Institute, Zhejiang Lab, Hangzhou, China

ARTICLE INFO

Article history:
Received 6 December 2021
Revised 2 November 2022
Accepted 24 November 2022
Available online 2 December 2022

Keywords:
Cross-modal hashing
semi-supervised learning
graph convolutional networks
modality-specific features
modality-shared features

ABSTRACT

Cross-modal hashing maps heterogeneous multimedia data into Hamming space for retrieving relevant samples across modalities, which has received great research interests due to its rapid retrieval and low storage cost. In real-world applications, due to high manual annotation cost of multi-media data, we can only make use of limited number of labeled data with rich unlabeled data. In recent years, several semi-supervised cross-modal hashing (SCH) methods have been presented. However, how to fully explore and jointly utilize the modality-specific (complementarity) and modality-shared (correlation) information for retrieval has not been well studied for existing SCH works. In this paper, we propose a novel SCH approach named Modality-specific and Cross-modal Graph Convolutional Networks (MCGCN). The network architecture contains two modality-specific channels and a cross-modal channel to learn modality-specific and shared representations for each modality, respectively. Graph convolutional network (GCN) is leveraged in these three channels to explore intra-modal and inter-modal similarity, and perform semantic information propagation from labeled data to unlabeled data. Modality-specific and shared representations for each modality are fused with attention scheme. To further reduce the modality gap, a discriminative model is designed, learning to classify the modality of representations, and network training is guided by adversarial scheme. Experiments on two widely used multi-modal datasets demonstrate MCGCN outperforms state-of-the-art semi-supervised/supervised cross-modal hashing methods.

1. Introduction

With the rapid growth of multi-media data, cross-modal retrieval [1–5] has received continuous research attention, whose goal is to search semantically relevant instances from one modality with the query instance of another modality [6,7]. One of the most popular pipeline is cross-modal hashing [8,9], which learns to convert multi-media data into binary hash codes for retrieval, due to its advantage in retrieval speed and storage for large-scale data [10,11]. Different modalities usually have inconsistent distributions and representations, which is the main challenge. To deal with this modality gap, several supervised cross-modal hashing methods have been developed [12], e.g., collective matrix factorization hashing (CMFH) [13], deep cross-modal hashing (DCMH) [8], cycle-consistent deep generative hashing (CYC-DGH) [14], etc.

Although supervised cross-modal hashing methods have achieved significant progress, they heavily rely on the semantic label information. However, labeling a large repository of instances containing multiple modalities is time and labor consuming and is infeasible. Some unsupervised cross-modal hashing methods have demonstrated that unlabeled multi-media data is also useful for the retrieval task [15,16]. For example, cluster-wise unsupervised hashing (CUH) [17] adopts the multi-view clustering manner to project data of different modalities into latent space to seek cluster centroid points for learning compact hash codes and linear hash functions. Focusing on the unsupervised retrieval task, aggregation-based graph convolutional hashing (AGCH) [18] uses multiple metrics to formulate affinity matrix for hash code learn-

* Corresponding authors.
E-mail address: xiaoyuanjing@whu.edu.cn (X.-Y. Jing).

https://doi.org/10.1016/j.patcog.2022.109211
0031-3203/

ing. Deep graph-neighbor coherence preserving network (DGCPN) [19] presents graph-neighbor coherence to explore the relationships between unlabeled data and its neighbors, and adopts a comprehensive similarity preserving loss for preserving similarity.

In real-world application, we usually can obtain a small quantity of labeled multi-media data and access rich unlabeled data with multiple modalities to perform cross-modal hashing in this semi-supervised scenario. In recent years, benefited from the development of deep learning technology [20], a few deep learning based semi-supervised cross-modal hashing (SCH) methods have been presented and demonstrated to bring favorable retrieval performance, e.g., semi-supervised deep quantization (SSDQ) [21], ranking-based deep cross-modal hashing (RDCMH) [22], semi-supervised cross-modal hashing approach by generative adversarial network (SCH-GAN) [23], etc. Recently, the powerful representation learning technology, i.e., graph convolutional network (GCN) [24], has been successfully introduced into SCH [25]. Semi-supervised graph convolutional hashing network (SGCH) [26] preserves high-order intra-modality similarity with GCN and adopts a siamese network to map the learned node representations into hamming space for achieving hash codes.

1.1. Motivation and contribution

Although a set of SCH methods have been developed, existing SCH methods mainly focus on intra-modal feature learning and similarity preserving, and then build bridge across modalities in the way of loss function establishment, e.g., [21,22] and [25], or a certain network module, e.g., [23] and [26], with the learned features of each modality for reducing the modality gap and learning hash codes. How to jointly explore both intra-modal and inter-modal semantic similarity and structure information in labeled and unlabeled data, such that the modality-specific and modality-shared information is fully exploited and used, has not been well studied. In this paper, we propose a novel SCH approach named Modality-specific and Cross-modal Graph Convolutional Networks (MCGCN). The contributions of our work are summarized as following three points:

(1) MCGCN provides a three-channel network architecture, including two modality-specific channels and a cross-modal channel for image and text modalities. Besides intra-modal graph modeling, cross-modal graph is also modeled with heterogeneous image and text features. Joint intra- and inter-modal semantic similarity preservation and semantic information propagation for unlabeled samples are performed based on GCN. And the modality-specific and shared representations are fused with attention scheme for each modality. To our knowledge, this is the first work to specially build cross-modal graph and jointly learn modality-specific and modality-shared features for SCH.
(2) The adversarial scheme is employed to guide optimization of network parameters. The generative model learns to predict the semantic labels of feature representations, and makes full use of the label and semantic similarity information to generate discriminant hash codes. And the discriminative model builds modality classifier to model inter-modal invariance with the adversarial loss.
(3) We evaluate MCGCN on the widely used benchmark datasets Wikipedia [27] and NUS-WIDE-10K [28]. The experimental results demonstrate our approach can achieve state-of-the-art SCH performance.

1.2. Organization

The rest of this paper is organized as follows. Section 2 briefly introduces the related works on supervised and unsupervised cross-modal hashing methods, semi-supervised cross-modal hashing methods, and graph convolutional networks. In Section 3, we detail the proposed MCGCN approach. Section 4 reports the experimental results on the Wikipedia and NUS-WIDE-10K datasets, and provides a comprehensive discussion about MCGCN. Finally, the conclusions are drawn in Section 5.

2. Related works

2.1. Supervised and unsupervised cross-modal hashing methods

Nowadays, several supervised or unsupervised cross-modal hashing methods have been presented and have achieved significant process [29–32]. With the matrix factorization technology, collective matrix factorization hashing (CMFH) [13] tries to learn unified hash codes in the shared latent semantic space for different modalities of an instance. Deep cross-modal hashing (DCMH) [8] provides an end-to-end deep learning framework to perform cross-modal retrieval. Cycle-consistent deep generative hashing (CYC-DGH) [14] adopts the adversarial training scheme to learn a couple of hash functions that can realize translation between modalities for reducing the heterogeneity. Robust and discrete matrix factorization hashing (RDMH) [33] learns the binary codes without any relaxation, avoiding the quantization loss, and uses the semantic label embedding scheme to find the relationships between semantic labels and hash codes.

The unsupervised generative adversarial cross-modal hashing (UGACH) method [34] tries to use the ability of unsupervised representation learning of generative adversarial network (GAN) to exploit the manifold structure in data of different modalities. Unsupervised coupled cycle generative adversarial hashing networks (UCH) [35] adopts the outer-cycle network to learn common representations, and uses the inner-cycle network to generate reliable hash codes. Deep graph-neighbor coherence preserving network (DGCPN) [19] tries to exploit similarity, i.e., the graph-neighbor coherence, the coexistent similarity, and the intra- and inter-modality consistency, in unlabeled multi-modality data.

However, these methods cannot be directly used in the semi-supervised scenario, where there exist both labeled multi-media data and rich unlabeled data with multiple modalities.

2.2. Semi-supervised cross-modal hashing methods

In recent years, to be flexible to use both labeled and unlabeled data from multiple modalities, a few semi-supervised cross-modal hashing (SCH) methods have been developed. The semi-supervised semantic-preserving hashing (S3PH) method [36] integrates the relaxed latent subspace learning and semantic-preserving regularization into a unified optimization objective. Focusing on composite quantization, the semi-supervised deep quantization (SSDQ) method [21] incorporates the information of paired data, labeled data and unlabeled data into a single framework. Focusing on the issue of incomplete and insufficient labels of multi-media data, the weakly-supervised cross-modal hashing (WCHash) method [37] enriches the labels of training data for learning cross-modal hashing functions. The ranking-based deep cross-modal hashing (RDCMH) method [22] learns the semi-supervised semantic ranking list based on the feature and label information of data, and then integrates the semantic ranking information into the deep cross-modal hashing process. The semi-supervised cross-modal hashing approach by generative adversarial network (SCH-GAN) method [23] provides a GAN-based solution for cross-modal hashing and uses reinforcement learning for optimization, where the generative model learns to select margin examples for the given cross-modal query and the discriminative model tries to judge their relevance.

There exist obvious differences between our approach and these SCH methods. Our approach utilizes the graph neural network to exploit relationship between samples and conduct semantic information propagation, and models both intra-modal and cross-modal graphs to jointly explore the modality-specific and modality-shared information for learning discriminative hash codes.

2.3. Graph convolutional networks

Graph convolutional network (GCN) presented by Kipf and Welling [24] has been demonstrated to be effective for semi-supervised classification [38–40]. It provides an efficient layer-wise propagation rule that can directly perform convolution operation on graphs, which is a powerful technology for analyzing graph data. Its main idea is to aggregate information from local graph neighborhoods and perform semi-supervised learning based on the fact that similar connected nodes have a large probability to be from the same class. Recently, GCN has been successfully introduced into the cross-modal retrieval task. Xu et al. [41] presented the graph convolutional hashing (GCH) method, which uses a semantic encoder for semantic information exploiting and adopts GCN to explore the similarity structure among nodes for generating favorable hash codes. Duan et al. [25] developed the semi-supervised cross-modal graph convolutional network hashing (CMGCNH) method, which applies asymmetric GCN for retrieval, and performs cooperative multimodal learning to learn hash codes. Shen et al. [26] presented the semi-supervised graph convolutional hashing network (SGCH) method, which preserves intra-modal similarity with GCN and tries to realize distribution agreement across modalities with the adversarial loss.

Different from these GCN-based cross-modal retrieval methods, we for the first time focus on cross-modal graph modeling and representation learning, and we provide a joint intra- and inter-modal graph structure exploration solution that achieves the state-of-the-art retrieval performance.

3. Proposed approach

3.1. Notation

Given multimodal dataset $D = \{I, T\}$, where $I = [i_1, \ldots, i_N] \in \mathbb{R}^{d_I \times N}$ and $T = [t_1, \ldots, t_N] \in \mathbb{R}^{d_T \times N}$ separately denote the feature matrices for the image and text modalities, which can be divided into a retrieval set D_r and a query set D_q. Here, N is the total number of feature vectors of image/text modalities and $d_I \neq d_T$. The retrieval set $D_r = \left\{D_r^L, D_r^U\right\}$, where D_r^L is a collection of N_L instances of labeled image-text pairs, and D_r^U is a set of N_{RU} instances of unlabeled image-text pairs. $l_p^L \in \{0,1\}^{C\times 1}$ represents the class label vector of the pth instance $o_p^L = \left(i_p^L, t_p^L\right)$ in D_r^L, where i_p^L and t_p^L separately denote the labeled image and text feature vectors, and C denotes the total number of classes. If o_p^L is from the cth class, $l_{pc}^L = 1$, otherwise, $l_{pc}^L = 0$. The query set D_q includes N_Q unlabeled pairs of image features and text features. In this paper, we employ graph convolution networks (GCN) to explore the structure information of labeled and unlabeled samples in graph. For convenience, we denote the total set of unlabeled image-text pairs as $D^U = \left\{o_p^U\right\}_{p=1}^{N_U} = \left\{\left(i_p^U, t_p^U\right)\right\}_{p=1}^{N_U}$, where N_U denotes the total number of unlabeled instances. The objective of cross-modal hashing is to learn hash functions for generating discriminative hash code of each input image/text feature vector.

3.2. Network architecture

The overall architecture of our approach MCGCN is shown in Fig. 1. It consists of three modules, i.e., intra-modal and cross-modal graph modeling, graph convolutional representation learning, and adversarial learning. In the graph construction module, we separately define the intra-modal adjacency matrices G^I and G^T for image and text modalities, and employ auto-encoders to obtain encoded representations with the same dimensionality for I and T for constructing cross-modal adjacency matrix G^S. The graph convolutional representation learning module uses GCN to explore intra-modal and inter-modal semantic similarity for obtaining discriminant feature representations, and utilizes attention scheme to fuse modality-specific and shared representations. In the adversarial learning module, the generative model learns to predict the semantic labels of features and learns to generate discriminative hash codes, and the discriminative model learns to classify the modality of features.

3.3. Intra-modal and cross-modal graph modeling

To make use of the powerful representation learning ability of GCN, we need to explore discriminant information on the graph data. To fully explore intra-modal semantic similarity, we separately build undirected graphs $\mathcal{G}_* = (\nu_*, \varepsilon_*), * \in \{I, T\}$, which denote the graphs of size N with nodes $x_n = i_n(t_n) \in \nu_*$ and edges $(x_n, x_m) \in \varepsilon_*$, for image and text modalities. With the established graph $\mathcal{G}_I$, the adjacency matrix G^I for the image modality is defined as

$$G_{mn}^I = \begin{cases} 1, & \text{if } i_m \text{ and } i_n \text{ are labeled and } l_m = l_n \\ 1, & \text{if } i_m \text{ or } i_n \text{ is unlabeled and } i_m(i_n) \in \mathcal{N}_r(i_n(i_m)) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where l_n is the one-hot class label vector of i_n, $i_n \in \mathcal{N}_r(i_m)$ denotes that i_n belongs to the r nearest neighbors of i_m. The adjacency matrix G^T for the text modality can be defined in the same manner. In this paper, r is empirically set as $r = 20$.

Besides modality-specific semantic similarity exploration, we also make effort to explore inter-modal semantic similarity through cross-modal graph modeling. However, the dimensions of feature vectors from different modalities are usually different, such that the features across modalities can not be directly compared. Inspired by the idea of dimensionality reduction and feature reconstruction of autoencoder [42], in this paper, we introduce autoencoders to obtain latent representations with the same dimensionality for different modalities. Specifically, we adopt the autoencoder with one fully connected layer for the encoder and decoder parts for each modality. Given the feature matrices for the image and text modalities I and T, the encoders learn latent representations $I_e = f_e^I(I) \in \mathbb{R}^{d\times N}$ and $T_e = f_e^T(T) \in \mathbb{R}^{d\times N}$ with mapping functions f_e^I and f_e^T for two modalities, and the decoders map the representations back to the reconstruction $I_d = f_d^I(I_e)$ and $T_d = f_d^T(T_e)$ with the mapping functions f_d^I and f_d^T. θ_{ae}^I and θ_{ae}^T are the parameters of the autoencoders for the image and text modalities. We should minimize the reconstruction loss $L_r(\theta_{ae}^I, \theta_{ae}^T)$ with the objective shown in Fig. 2.

We build the cross-modal graph as $\mathcal{G}_S = (\nu_S, \varepsilon_S)$, where ν_S is the vertex set corresponding to the total representation set $S_{IT} = \{I_e, T_e\}$ and ε_S is the collection of edges. We define the cross-modal adjacency matrix G^S based on intra-modal adjacency matrices G^I and G^T, since we deem that if any two image(text) features, e.g., $i_m(t_m)$ and $i_n(t_n)$ are connected, the corresponding encoder outputs $i_e^m(t_e^m)$ and $i_e^n(t_e^n)$ should be connected, and $i_e^m(t_e^m)$ and $t_e^n(i_e^n)$ should also be connected, leaving aside the modality difference. Specifically, G^S is defined as follows

$$G^S = \begin{bmatrix} G^I & \frac{G^I+G^T}{2} \\ \frac{G^I+G^T}{2} & G^T \end{bmatrix} \tag{2}$$

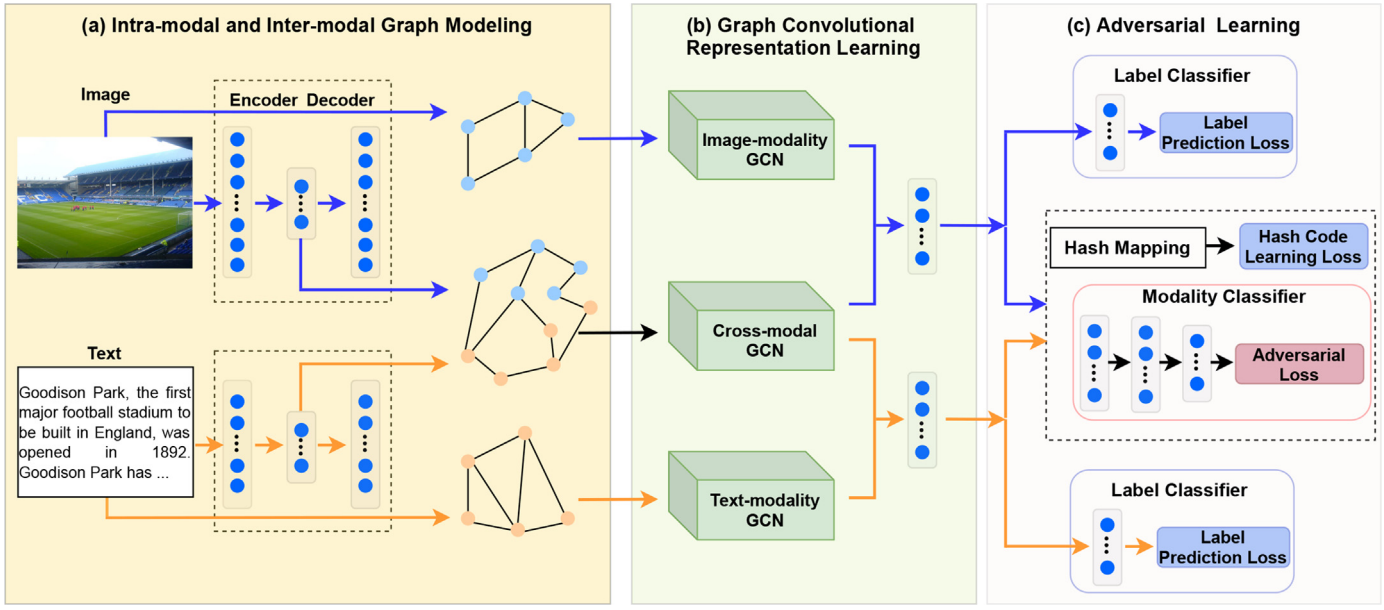

Fig. 1. The overall framework of our MCGCN approach.

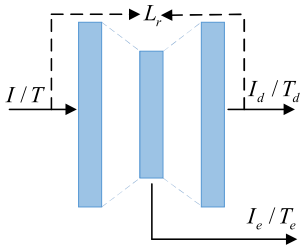

Fig. 2. The objective of the reconstruction loss in the autoencoders of two modalities.

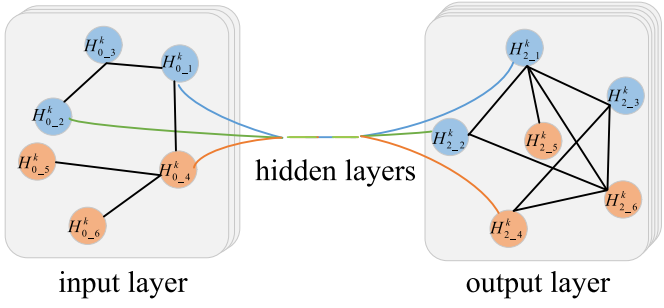

Fig. 3. Schematic depiction of GCN.

3.4. Graph convolutional representation learning

With the established modality-specific graphs $\mathcal{G}_I$ and $\mathcal{G}_T$, and the cross-modal graph $\mathcal{G}_S$, we build a three-channel network module based on graph convolutional networks (GCN) to learn modality-specific and shared representations for each modality. Specifically, for each channel, a two-layer GCN is used, and the layer-wise propagation $H^k_{l+1} = f^k_l\left(H^k_l, G^k; \Theta^k_l\right)$, $k = \{I, T, S\}$ is performed for the lth layer of GCN. The detailed process is shown in Fig. 3.

The output feature representation set H^k_{l+1} can be obtained through graph convolutional function f^k_l based on input feature set H^k_l and the adjacency matrix G^k. $l = 0, 1$ and $\Theta^k = \left\{\Theta^k_0, \Theta^k_1\right\}$ is the set of parameters of two layers. H^k_0 is the input of GCN, i.e., $H^{I(T)}_0 = I(T)$ for the image and text modalities, and $H^S_0 = S_{IT}$ for the cross-modal channel. The convolution function in [24] is used

$$f^k_l = \tanh\left(\left(D^k\right)^{-\frac{1}{2}}\tilde{G}^k\left(D^k\right)^{-\frac{1}{2}}H^k_l\Theta^k_l\right) \tag{3}$$

where $\tilde{G}^k = G^k + \mathbf{I}$, $\mathbf{I}$ is an identity matrix, D^k is a degree matrix with the diagonal element $D^k_{nn} = \sum_m \tilde{G}^k_{mn}$, and $\tanh(\cdot)$ is an activation function. In this way, GCN can explore and preserve the high-order intra-modal and inter-modal similarity, and perform semantic information propagation from unlabeled samples to labeled samples.

When we obtain the output feature representations of GCN corresponding to the image-modality, text-modality, and cross-modal channels, i.e., $Z^I = H^I_2$, $Z^T = H^T_2$, and $\left[Z^I_S, Z^T_S\right] = H^S_2$, we fuse the modality-specific and modality-shared representations to generate unified representations for each specific modality, such that subsequent hashing learning can be performed based on the intra-modal characteristics and the correlation (commonness) across modalities jointly for each modality. Considering that the modality-specific representations and modality-shared representations may contribute differently to the unified representations, we provide a cross-graph attention scheme to seek the significance of each kind of representations for fusion. Specifically, we adopt an attention function $f^I_a\left(\cdot; \theta^I_a\right)$, i.e., a single-layer feed-forward sub-network activated by the Sigmoid function and parameterized by θ^I_a, to obtain the attention coefficient for the image modality

$$\begin{cases} A^I = f^I_a\left(Z^I; \theta^I_a\right) \\ A^I_S = f^I_a\left(Z^I_S; \theta^I_a\right) \end{cases} \tag{4}$$

The Softmax function is used to further normalize these coefficients. Then, we can obtain the fused node feature representation set R^I for the image modality

$$R^I = A^I Z^I + A^I_S Z^I_S \tag{5}$$

For the text modality, the fused feature representation set R^T can be obtained in the same way.

3.5. Adversarial learning

In this section, we will introduce the loss functions of the proposed approach, including the label prediction loss, hash code

learning loss, and the adversarial loss. And network training will be guided by the adversarial mechanism between the generative model and the discriminative model.

3.5.1. Generative model

To further make the fused feature representations be semantically discriminative, we build a mapping from the feature representation space to the semantic label space. We adopt a one-layer sub-network activated by the Softmax function as a classifier for each modality. When the labeled representation of $R^l(R^T)$, i.e., r_p^l or r_p^T, is input into the corresponding classifier, the probability distribution of semantic categories of the feature representation, i.e., $P^l(r_p^l)$ or $P^T(r_p^T)$, can be obtained. We utilize these probability distributions to define the label prediction loss as follows

$$L_{lp}(\theta_{lp}^l, \theta_{lp}^T) = -\frac{1}{N_L}\sum_{p=1}^{N_L} l_p^L\left(\log P^l(r_p^l) + \log P^T(r_p^T)\right) \tag{6}$$

To facilitate efficient retrieval with significantly reduced storage needs, we map feature representations into Hamming space to obtain the corresponding hash codes. Specifically, the hash codes can be obtained through $B^* = sign(R^*) \in \{-1, +1\}^{v\times N}$, $* \in \{I, T\}$, where $sign(\cdot)$ is the element-wise sign function. Each column of B^* is the learned v-bit hash codes. Cross-modal semantic similarity is expected to be preserved between feature representations and between the corresponding hash codes. Inspired by Jiang and Li [8], Xu et al. [41], we provide the following hash code learning loss with semantic similarity preservation

$$\begin{aligned} L_{hcl}(\Theta^I, \Theta^T, \Theta^S, \theta_a^I, \theta_a^T) = &-\sum_{p,q=1}^{N_L}\left(J_{pq}\omega_{pq} - \log\left(1 + e^{\omega_{pq}}\right)\right) \\ &+\alpha\left(\left\|B^I - R^I\right\|_F^2 + \left\|B^T - R^T\right\|_F^2\right) + \beta\left(\left\|B^I\mathbf{1}\right\|_F^2 + \left\|B^T\mathbf{1}\right\|_F^2\right) \end{aligned} \tag{7}$$

where $\omega_{pq} = \frac{1}{2}r_p^{I\prime}r_q^T$, $\mathbf{1}$ is a vector with all elements being 1, α and β are balance factors, and $(\cdot)'$ denotes the transposition operation. J is the cross-modal semantic similarity matrix, and $J_{pq} = 1$, if r_p^I and r_q^T is from the same class, and $J_{pq} = 0$, otherwise. The first term is a cross-modal semantic similarity preservation loss on features, which is the negative log-likelihood function. By minimizing this term, the similarity of feature representations of the same class across modalities will be maximized, while the cross-modal similarity of representations from different classes will be minimized simultaneously. The second term is the approximation loss for the fused feature representations and corresponding hash codes. Through this approximation, the hash codes are also excepted to preserve the cross-modal semantic similarity. The third term is used to promote each bit of hash code to be balanced for all input samples.

3.5.2. Discriminative model

To further reduce the modality gap, we build a modality classifier acting as an adversary, which aims to recognize the modality of fused feature representations, with a three-layer fully connected sub-network. The adversarial loss is defined as follows

$$L_{adv}(\theta_A) = -\frac{1}{N}\sum_{n=1}^{N}\left(\log A(r_n^I; \theta_A) + \log\left(1 - A(r_n^T; \theta_A)\right)\right) \tag{8}$$

where $A(r_n^I; \theta_A)$ denotes the probability of modality for the representation r_n^I, and θ_A is the parameter of the sub-network. By defining this cross-entropy based loss, we intend to reduce the cross-modal heterogeneity gap in the level of features with the adversarial scheme.

3.5.3. Optimization

We should jointly minimize the hash code learning loss L_{hcl} in Eq. (7), the label prediction loss L_{lp} in Eq. (6), and the reconstruction loss L_r in Fig. 2 of the generative model, and minimize the adversarial loss L_{adv} in Eq. (8) of the discriminative model. Considering that the generative model and discriminative model have opposite optimization goals, we use mini-max game for optimization

$$\begin{aligned} \left(\hat{\theta}_{ae}^*, \hat{\Theta}^*, \hat{\Theta}^S, \hat{\theta}_a^*, \hat{\theta}_{lp}^*\right) = \arg\min_{\theta_{ae}^*, \Theta^*, \Theta^S, \theta_a^*, \theta_{lp}^*} & L_{hcl}(\Theta^*, \Theta^S, \theta_a^*) \\ &+ \gamma L_{lp}(\theta_{lp}^*) + \eta L_r(\theta_{ae}^*) - L_{adv}\left(\hat{\theta}_A\right) \end{aligned} \tag{9}$$

$$\begin{aligned} \left(\hat{\theta}_A\right) = \arg\max_{\theta_A} & L_{hcl}\left(\hat{\Theta}^*, \hat{\Theta}^S, \hat{\theta}_a^*\right) + \gamma L_{lp}\left(\hat{\theta}_{lp}^*\right) \\ &+ \eta L_r\left(\hat{\theta}_{ae}^*\right) - L_{adv}(\theta_A) \end{aligned} \tag{10}$$

where γ and η are hyper-parameters to balance three terms of the generative model, and $* = \{I, T\}$. These parameters are updated by using the stochastic gradient descent algorithm. Following [43], a gradient reversal layer (GRL) is added before the first layer of the modality classifier to facilitate optimization. The optimization process is briefly summarized in Algorithm 1.

Algorithm 1 Optimization of MCGCN.

1. **Input**: Image and text features I and T, and class label set $\{l_p^L\}_{p=1}^{N_L}$ of labeled feature set.
2. **Construct intra-modal and cross-modal graphs** $\mathcal{G}_I$, $\mathcal{G}_T$, and $\mathcal{G}_S$.
3. **Update until convergence**
   (a) Separately update $\theta_{ae}^*, \Theta^*, \Theta^S, \theta_a^*, \theta_{lp}^*$, $* = \{I, T\}$ by descending their stochastic gradients with the learning rate ρ:
   $\theta_{ae}^* \leftarrow \theta_{ae}^* - \rho\nabla_{\theta_{ae}^*}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$,
   $\Theta^* \leftarrow \Theta^* - \rho\nabla_{\Theta^*}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$,
   $\Theta^S \leftarrow \Theta^S - \rho\nabla_{\Theta^S}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$,
   $\theta_a^* \leftarrow \theta_a^* - \rho\nabla_{\theta_a^*}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$,
   $\theta_{lp}^* \leftarrow \theta_{lp}^* - \rho\nabla_{\theta_{lp}^*}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$.
   (b) Update θ_A by ascending the stochastic gradients through GRL:
   $\theta_A \leftarrow \theta_A + \rho\nabla_{\theta_A}\frac{1}{N}\left(L_{hcl} + \gamma L_{lp} + \eta L_r - L_{adv}\right)$.
4. **Output**: Hash codes B^I and B^T of the image and text modalities.

4. Experiments

4.1. Datasets and compared methods

In this paper, we use two benchmark datasets Wikipedia [27] and NUS-WIDE-10K [28] to evaluate our approach MCGCN.

-The Wikipedia dataset [27] is collected from Wikipedia articles. It contains 2,866 image-text pairs from 10 categories. Following [23], the dataset is divided into a training set (retrieval set) with 2,173 pairs and a test set (query set) with the remaining 693 pairs.

-The NUS-WIDE-10K dataset [28] is a subset of the NUS-WIDE dataset [44], including the pairs of 10 largest categories of NUS-WIDE. It contains 10,000 image-text pairs, where 8,000 pairs are selected to constitute the training set (retrieval set), and the remaining 2,000 pairs are used for testing (query set).

On these two datasets, following [26,43,45], 4,096d feature vectors extracted by the Fc7 layer of the VGGNet are used to represent images. The 3,000d bag-of-words (BoW) feature vectors are used to represent texts for Wikipedia, and 1,000d BoW vectors are used for text features on NUS-WIDE-10K.

Six state-of-the-art and related cross-modal retrieval methods are used as baselines for comparison, including three semi-supervised cross-modal hashing methods, i.e., RDCMH [22], SGCH [26], SCH-GAN [23], an unsupervised cross-modal hashing method, i.e., DGCPN [19], a GCN-based supervised cross-modal hashing method, i.e., GCH [41], and a non-hashing semi-supervised cross-modal retrieval method, i.e., SMLN [4]. For semi-supervised and unsupervised methods, all labeled and unlabeled available data is used for training, while only the labeled data is used for training for the supervised method GCH. For these compared methods, the source codes are kindly provided by the authors. For fairness, we use the same experimental setting in this paper for these baselines for experiment. We carefully tune the hyper-parameters as recommended by the original papers.

4.2. Implementation details and evaluation measures

The details of the network are as follows: we deploy the autoencoder for each modality to learn latent representations, i.e., $d_I(d_T) \to 2048 \to d_I(d_T)$ for Wikipedia and $d_I(d_T) \to 1024 \to d_I(d_T)$ for NUS-WIDE-10K, where the ReLU activation function is used after each fully connected layer. In the graph convolutional representation learning module, a two-layer GCN based three-channel sub-network is used, where the image-modality GCN performs $d_I \to 512 \to v$ for Wikipedia and $d_I \to 1024 \to v$ for NUS-WIDE-10K, the text-modality GCN performs $d_T \to 512 \to v$ for both datasets, and the cross-modal GCN performs $2048 \to 1024 \to v$ for Wikipedia and $1024 \to 512 \to v$ for NUS-WIDE-10K. For modality classification, three layers (i.e., $v \to 16 \to 8 \to 2$) activated by the ReLU function are used for both datasets. The learning rate is set as 0.0001.

In this paper, we tune the hyper-parameters (balance factors α and β in Eq. (7), and γ and η in Eq. (9)) using the grid search strategy. The search range of α, β and η is $\left[10^{-3}, 10^{2}\right]$ and the range of γ is $\left[10^{1}, 10^{5}\right]$, with 10 times per step. Specifically, these parameters are set as: $\beta = 0.1$, $\gamma = 1000$, $\eta = 0.01$ for both datasets, $\alpha = 1$ for Wikipedia and $\alpha = 0.1$ for NUS-WIDE-10K.

In this paper, we randomly select 30% and 70% image-text pairs in the training set as labeled data, and mask the labels of the remaining pairs as the unlabeled data. Following [22], we also report the retrieval results with 100% training data being used for the labeled data. We investigate two cross-modal retrieval tasks, i.e., retrieving text given an image query (I2T) and image retrieval using a text query (T2I). The retrieval performance is evaluated by using mean average precision (MAP) [45]. To evaluate the influence of random running and random partition for the labeled and unlabeled data, we perform 10 random runnings to report the average results across 10 random runnings (partitions).

4.3. Comparison with state-of-the-arts

Tables 1 and 2 separately show the cross-modal retrieval results (average MAP result (± standard deviation)) on MAP of our MCGCN and other compared methods on Wikipedia and NUS-WIDE-10K datasets. It is noted that in the tables, 2.9E-3 means $2.9 \times 10^{-3} = 0.0029$. The best results are highlighted in bold. From the tables, we have the following observations: (1) As the size of labeled data and the length of hash codes increase, better retrieval performances will be achieved for compared methods in most cases. DGCPN is an unsupervised method that uses all training data as unlabeled data. Thus, it seems that it is not sensitive to the size of

Table 1
MAP results (average result (± standard deviation)) of compared methods with various rates of labeled samples in the training set on Wikipedia.

Task	Method	16-bit			32-bit			64-bit		
		30%	70%	100%	30%	70%	100%	30%	70%	100%
I2T	DGCPN [19]	0.404 ± 2.9E − 3	0.404 ± 2.9E − 3	0.404 ± 2.9E − 3	0.411 ± 3.0E − 3	0.411 ± 3.0E − 3	0.411 ± 3.0E − 3	0.420 ± 9.0E − 3	0.420 ± 9.0E − 3	0.420 ± 9.0E − 3
	GCH [41]	0.326 ± 6.0E − 3	0.432 ± 1.5E − 2	0.568 ± 4.7E − 3	0.388 ± 8.2E − 3	0.482 ± 4.4E − 3	0.615 ± 4.0E − 3	0.418 ± 1.5E − 2	0.506 ± 1.3E − 2	0.632 ± 6.3E − 3
	SMLN [4]	0.449 ± 1.8E − 2	0.566 ± 1.0E − 2	0.676 ± 1.8E − 2	0.449 ± 1.8E − 2	0.566 ± 1.0E − 2	0.676 ± 1.8E − 2	0.449 ± 1.8E − 2	0.566 ± 1.0E − 2	0.676 ± 1.8E − 2
	RDCMH [22]	0.334 ± 2.5E − 2	0.423 ± 1.2E − 2	0.461 ± 2.0E − 2	0.356 ± 7.2E − 3	0.437 ± 4.2E − 3	0.497 ± 1.7E − 2	0.367 ± 9.5E − 2	0.462 ± 1.1E − 2	0.500 ± 2.2E − 2
	SGCH [26]	0.452 ± 3.9E − 2	0.572 ± 2.1E − 2	0.747 ± 5.8E − 4	0.509 ± 4.7E − 3	0.675 ± 8.6E − 3	0.784 ± 1.5E − 3	0.524 ± 1.0E − 2	0.680 ± 4.0E − 3	0.795 ± 1.2E − 3
	SCH-GAN [23]	0.427 ± 1.9E − 2	0.472 ± 2.2E − 2	0.527 ± 5.7E − 3	0.429 ± 1.4E − 2	0.477 ± 3.4E − 3	0.528 ± 1.2E − 2	0.458 ± 6.0E − 3	0.491 ± 1.2E − 2	0.565 ± 5.6E − 3
	MCGCN	**0.544** ± 7.0E − 3	**0.678** ± 5.0E − 3	**0.770** ± 2.1E − 3	**0.638** ± 6.3E − 3	**0.747** ± 7.0E − 3	**0.824** ± 3.6E − 3	**0.654** ± 7.6E − 3	**0.751** ± 2.1E − 3	**0.825** ± 1.0E − 3
T2I	DGCPN [19]	0.440 ± 2.0E − 3	0.440 ± 2.0E − 3	0.440 ± 2.0E − 3	0.474 ± 3.2E − 3	0.474 ± 3.2E − 3	0.474 ± 3.2E − 3	0.489 ± 7.5E − 3	0.489 ± 7.5E − 3	0.489 ± 7.5E − 3
	GCH [41]	0.355 ± 4.2E − 3	0.579 ± 4.5E − 3	0.741 ± 5.5E − 3	0.413 ± 6.4E − 3	0.644 ± 4.5E − 3	0.752 ± 9.1E − 3	0.433 ± 1.3E − 2	0.675 ± 8.2E − 3	0.783 ± 3.1E − 3
	SMLN [4]	0.413 ± 1.5E − 2	0.560 ± 2.8E − 3	0.664 ± 3.4E − 3	0.413 ± 1.5E − 2	0.560 ± 2.8E − 3	0.664 ± 3.4E − 3	0.413 ± 1.5E − 2	0.560 ± 2.8E − 3	0.664 ± 3.4E − 3
	RDCMH [22]	0.359 ± 3.9E − 2	0.423 ± 1.2E − 2	0.450 ± 2.0E − 2	0.367 ± 7.2E − 3	0.440 ± 4.2E − 3	0.485 ± 1.7E − 2	0.384 ± 9.5E − 2	0.468 ± 1.1E − 2	0.515 ± 2.2E − 2
	SGCH [26]	0.463 ± 6.8E − 2	0.660 ± 2.1E − 2	0.758 ± 5.8E − 4	0.502 ± 2.3E − 3	0.710 ± 2.3E − 2	0.803 ± 2.2E − 2	0.513 ± 2.7E − 3	0.715 ± 1.3E − 2	0.815 ± 3.2E − 3
	SCH-GAN [23]	0.460 ± 1.7E − 2	0.668 ± 2.7E − 2	0.762 ± 1.8E − 2	0.508 ± 7.5E − 3	0.685 ± 1.8E − 2	0.810 ± 1.1E − 2	0.520 ± 7.6E − 3	0.690 ± 7.3E − 3	0.816 ± 8.2E − 3
	MCGCN	**0.553** ± 5.3E − 3	**0.680** ± 8.1E − 3	**0.765** ± 1.5E − 3	**0.641** ± 5.5E − 3	**0.745** ± 6.6E − 3	**0.823** ± 1.2E − 3	**0.653** ± 8.2E − 3	**0.751** ± 2.3E − 3	**0.824** ± 5.8E − 4

Table 2
MAP results (average result (± standard deviation)) of compared methods with various rates of labeled samples in the training set on NUS-WIDE-10K.

Task	Method	16-bit			32-bit			64-bit		
		30%	70%	100%	30%	70%	100%	30%	70%	100%
I2T	DGCPN [19]	$0.448 \pm 1.7E-2$	$0.448 \pm 1.7E-2$	$0.448 \pm 1.7E-2$	$0.454 \pm 1.6E-1$	$0.454 \pm 1.6E-1$	$0.454 \pm 1.6E-1$	$0.464 \pm 2.0E-3$	$0.464 \pm 2.0E-3$	$0.464 \pm 2.0E-3$
	GCH [41]	$0.444 \pm 2.4E-2$	$0.491 \pm 1.3E-2$	$0.611 \pm 3.6E-3$	$0.480 \pm 7.6E-3$	$0.509 \pm 1.2E-2$	$0.621 \pm 3.2E-3$	$0.499 \pm 1.1E-2$	$0.534 \pm 8.9E-3$	$0.662 \pm 2.9E-3$
	SMLN [4]	$0.502 \pm 1.2E-2$	$0.562 \pm 0.7E-3$	$0.705 \pm 2.3E-3$	$0.502 \pm 1.2E-2$	$0.562 \pm 0.7E-3$	$0.705 \pm 2.3E-3$	$0.502 \pm 1.2E-2$	$0.562 \pm 0.7E-3$	$0.705 \pm 2.3E-3$
	RDCMH [22]	$0.395 \pm 1.1E-2$	$0.498 \pm 5.7E-3$	$0.612 \pm 2.0E-2$	$0.399 \pm 2.1E-2$	$0.502 \pm 8.7E-3$	$0.620 \pm 1.0E-2$	$0.434 \pm 6.8E-3$	$0.527 \pm 3.3E-3$	$0.624 \pm 7.7E-2$
	SGCH [26]	$0.517 \pm 9.2E-3$	$0.586 \pm 1.4E-2$	$0.691 \pm 4.0E-3$	$0.553 \pm 7.2E-3$	$0.589 \pm 6.7E-3$	$0.730 \pm 1.8E-2$	$0.569 \pm 4.2E-3$	$0.659 \pm 5.3E-3$	$0.735 \pm 5.0E-3$
	SCH-GAN [23]	$0.520 \pm 2.3E-3$	$0.588 \pm 1.4E-2$	$0.713 \pm 3.1E-2$	$0.528 \pm 1.1E-2$	$0.589 \pm 1.4E-2$	$0.724 \pm 1.5E-2$	$0.550 \pm 1.8E-3$	$0.596 \pm 8.4E-3$	$0.733 \pm 6.0E-3$
	MCGCN	$\mathbf{0.569} \pm 4.5E-3$	$\mathbf{0.689} \pm 5.9E-3$	$\mathbf{0.753} \pm 3.6E-3$	$\mathbf{0.590} \pm 4.6E-3$	$\mathbf{0.701} \pm 2.5E-3$	$\mathbf{0.781} \pm 5.0E-3$	$\mathbf{0.594} \pm 4.0E-3$	$\mathbf{0.702} \pm 6.4E-3$	$\mathbf{0.782} \pm 5.0E-3$
T2I	DGCPN [19]	$0.467 \pm 8.2E-3$	$0.467 \pm 8.2E-3$	$0.467 \pm 8.2E-3$	$0.486 \pm 1.0E-3$	$0.486 \pm 1.0E-3$	$0.486 \pm 1.0E-3$	$0.491 \pm 1.2E-3$	$0.491 \pm 1.2E-3$	$0.491 \pm 1.2E-3$
	GCH [41]	$0.480 \pm 5.5E-3$	$0.648 \pm 1.0E-2$	$0.719 \pm 4.6E-3$	$0.531 \pm 1.1E-3$	$0.675 \pm 9.7E-3$	$0.763 \pm 5.1E-3$	$0.537 \pm 2.5E-3$	$0.688 \pm 2.7E-3$	$0.765 \pm 3.5E-3$
	SMLN [4]	$0.513 \pm 2.9E-2$	$0.597 \pm 1.7E-2$	$0.723 \pm 5.1E-3$	$0.513 \pm 2.9E-2$	$0.597 \pm 1.7E-2$	$0.723 \pm 5.1E-3$	$0.513 \pm 2.9E-2$	$0.597 \pm 1.7E-2$	$0.723 \pm 5.1E-3$
	RDCMH [22]	$0.407 \pm 2.5E-2$	$0.519 \pm 9.3E-2$	$0.667 \pm 1.7E-2$	$0.416 \pm 2.6E-2$	$0.523 \pm 6.9E-3$	$0.672 \pm 8.3E-2$	$0.455 \pm 3.7E-2$	$0.557 \pm 6.9E-3$	$0.678 \pm 7.1E-2$
	SGCH [26]	$0.521 \pm 3.5E-3$	$0.617 \pm 5.0E-3$	$0.717 \pm 2.6E-2$	$0.567 \pm 4.6E-3$	$0.620 \pm 5.0E-3$	$0.725 \pm 8.9E-3$	$0.571 \pm 2.5E-3$	$0.632 \pm 1.5E-3$	$0.743 \pm 5.0E-3$
	SCH-GAN [23]	$0.537 \pm 9.1E-3$	$0.672 \pm 1.6E-3$	$0.738 \pm 1.8E-2$	$0.546 \pm 3.2E-3$	$0.689 \pm 3.9E-3$	$0.743 \pm 3.3E-3$	$0.543 \pm 6.6E-3$	$0.693 \pm 2.9E-3$	$0.766 \pm 1.0E-3$
	MCGCN	$\mathbf{0.576} \pm 7.6E-3$	$\mathbf{0.687} \pm 6.7E-3$	$\mathbf{0.777} \pm 4.7E-3$	$\mathbf{0.592} \pm 4.5E-3$	$\mathbf{0.703} \pm 5.1E-3$	$\mathbf{0.780} \pm 2.6E-3$	$\mathbf{0.593} \pm 2.7E-3$	$\mathbf{0.705} \pm 5.9E-3$	$\mathbf{0.781} \pm 4.0E-3$

Table 3
Ablation study (on MAP) on the contribution of important components of MCGCN.

Task	Method	Wikipedia			NUS-WIDE-10K		
		16-bit	32-bit	64-bit	16-bit	32-bit	64-bit
I2T	MCGCN-S	0.521	0.569	0.575	0.539	0.564	0.572
	MCGCN-IT	0.468	0.495	0.497	0.504	0.510	0.524
	MCGCN-A	0.514	0.552	0.567	0.568	0.574	0.589
	MCGCN-C	0.477	0.588	0.589	0.504	0.567	0.568
	MCGCN	**0.544**	**0.638**	**0.654**	**0.569**	**0.590**	**0.594**
T2I	MCGCN-S	0.524	0.567	0.578	0.542	0.561	0.571
	MCGCN-IT	0.471	0.494	0.498	0.497	0.508	0.515
	MCGCN-A	0.513	0.550	0.568	0.566	0.571	0.591
	MCGCN-C	0.478	0.583	0.593	0.508	0.563	0.564
	MCGCN	**0.553**	**0.641**	**0.653**	**0.576**	**0.592**	**0.593**

labeled data. In addition, SMLN is a non-hashing semi-supervised cross-modal retrieval method. We list the same retrieval results of SMLN for different lengths of hash codes. (2) From these tables, we can see that our approach can always achieve the best retrieval results in different cases of the rates of labeled data and lengths of hash codes. Take labeled data rate of 30% and hash code length of 16 as an example, MCGCN improves MAP at least by 0.092=(0.544-0.452) in the case of I2T and 0.09=(0.553-0.463) in the case of T2I on Wikipedia, and by 0.049=(0.569-0.520) in the I2T case and 0.039=(0.576-0.537) in the T2I case on NUS-WIDE-10K. (3) Our approach achieves comparable standard deviation against competing methods.

The reasons of the performance improvement of our approach lie in the following three points: (a) By jointly intra-modal and cross-modal graph modeling and representation learning, both within-modal and between-modal structure and correlation information is well explored, such that modality-specific and modality-shared features are effectively fused and leveraged to generate hash codes for cross-modal retrieval. (b) The label and structure information of labeled and unlabeled samples are fully explored, and GCN is adopted to perform semantic information propagation. (c) The inter-modal invariance is elaborately modeled with the adversarial mechanism.

4.4. Discussion

4.4.1. Ablation study

In this subsection, we discuss the methodological details of MCGCN. We separately call the version of MCGCN without the cross-modal channel as MCGCN-S, the version of MCGCN without both the image-modality and text-modality channels as MCGCN-IT, the version of MCGCN without the cross-graph attention scheme as MCGCN-A, which concatenates the modality-specific and modality-shared representations for fusion, and the version of MCGCN without the modality classifier as MCGCN-C. Table 3 tabulates the MAP results of these variants of MCGCN when the rate of labeled data is 30%.

We can see that the MAP results of MCGCN-S and (especially) MCGCN-IT are obviously inferior to those of the complete version of MCGCN on two datasets. This phenomenon indicates that both effective intra-modal and cross-modal graph representation and learning are useful to the cross-modal retrieval task. In addition, results of MCGCN-A and MCGCN-C are also inferior to those of MCGCN, which means that the cross-graph attention based fusion and adversarial learning scheme with the modality classifier also contribute to the performance of our approach.

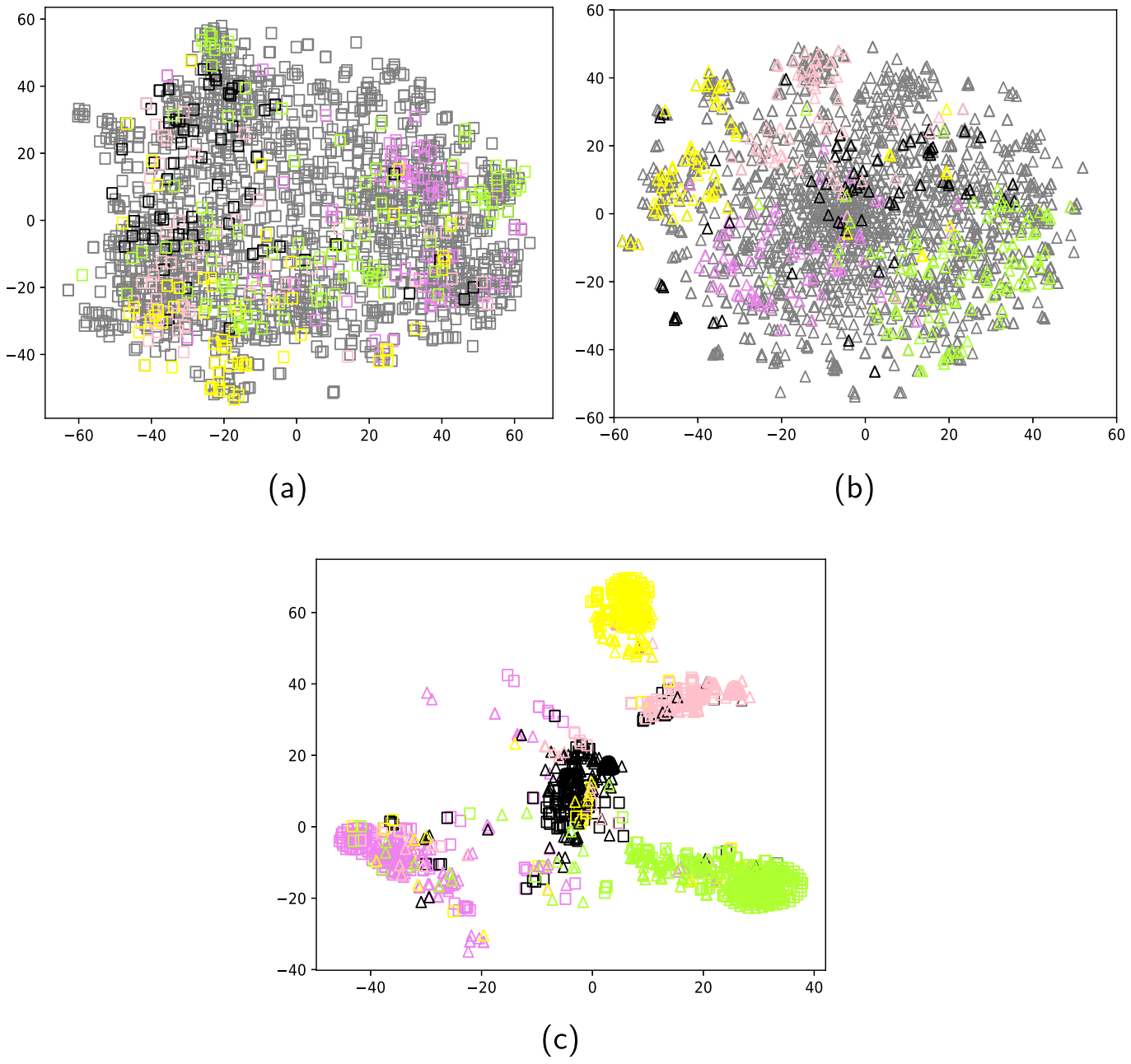

Fig. 4. T-SNE visualization of data on the Wikipedia dataset. In the figure, squares and triangles separately denote features/hash codes of image and text modalities. Different colors denote features/hash codes from different classes, and grayness denotes the unlabeled features.

4.4.2. Visualization of the learned representations

In order to further investigate the effectiveness of the learned representations by our MCGCN approach, we employ the t-SNE tool to embed the features/hash codes into the two-dimensional space for visualization. Taken the first five categories of training samples on the Wikipedia dataset (when 30% training samples are labeled) as an example, Fig. 4(a) and (b) separately show the distributions of original training samples (including labeled and unlabeled samples) from image and text modalities, and Fig. 4(c) shows the distributions of learned 32-bit hash codes of two modalities.

We can observe that the features with different class labels are not well separated and the distributions of two modalities are largely different in the original feature space. On the contrary, the learned hash codes from different classes are generally separated into ten semantically clusters. Furthermore, from Fig. 4(c), we can see that the distributions of image and text modalities are better mixed together. As a summary, this comparison indicates that our MCGCN approach can effectively reduce the modality gap and obtain hash codes with more favorable discriminant ability.

4.4.3. Parameter sensitivity

Lastly, we investigate the sensitivity of our approach to hyper-parameters α, β, γ and η. Figure 5 shows I2T/T2I retrieval results (on MAP) versus different values of α, β, γ and η on the Wikipedia dataset with 32-bit hash codes when the rate of labeled data is 30%. When one hyper-parameter is evaluated, the others are fixed. From the figure, MCGCN is not sensitive to the choice of α in the range of [0.001,1], β in the range of [0.01,0.1], and η in the range of [0.001,0.01]. The best results can be obtained when $\gamma = 1000$.

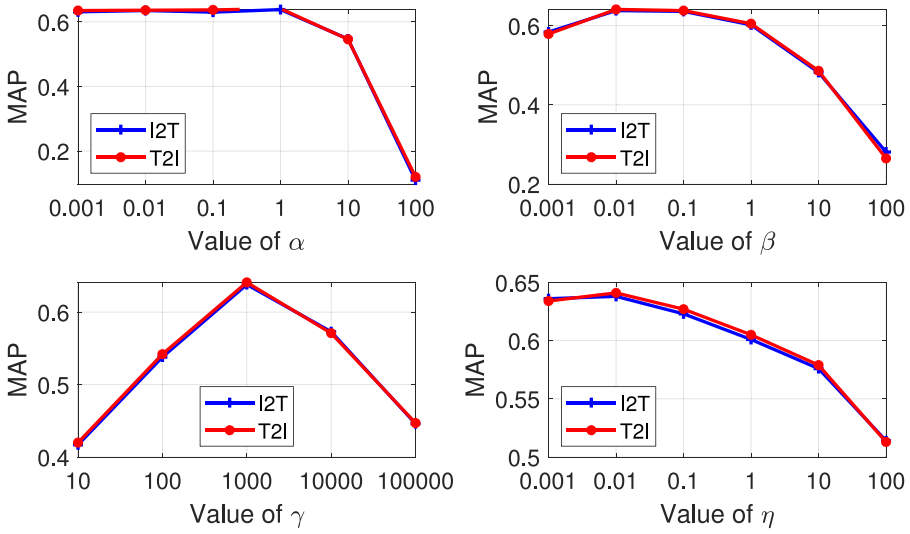

Fig. 5. Retrieval results of MCGCN versus different values of α, β, γ and η on Wikipedia.

For simplicity, these hyper-parameters are set as $\alpha = 1$, $\beta = 0.1$, $\gamma = 1000$ and $\eta = 0.01$ on Wikipedia. Similar experiment results can also be found on NUS-WIDE-10K.

5. Conclusion

In this paper, we propose a novel semi-supervised cross-modal hashing approach named MCGCN. Modality-specific and modality-shared features are effectively explored through joint intra-modal and cross-modal graph modeling and graph convolutional representation learning. The label and structure information of labeled and unlabeled samples are fully leveraged to perform semantic information propagation and learn discriminative hash codes.

Comprehensive experiments on two widely used datasets demonstrate that our approach performs better than state-of-the-art semi-supervised/supervised cross-modal retrieval methods. The experiment results also indicate the effectiveness of the adopted mechanisms in our approach, including modality-specific and cross-modal graph learning, cross-graph attention based fusion, and adversarial learning based optimization.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62076139, 61702280), Open Research Project of Zhejiang Lab (No. 2021KF0AB05), Future Network Scientific Research Fund Project (No. FNSRFP-2021-YB-15), 1311 Talent Program of Nanjing University of Posts and Telecommunications, the National Postdoctoral Program for Innovative Talents (No. BX20180146), China Postdoctoral Science Foundation (No. 2019M661901), and Jiangsu Planned Projects for Postdoctoral Research Funds (No. 2019K024).

References

[1] K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval, arXiv preprint arXiv:1607.06215 (2016).
[2] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, A. Torralba, Cross-modal scene networks, IEEE Trans. Pattern Anal. Mach.Intell. 40 (10) (2018) 2303–2314.
[3] D. Mandal, P. Rao, S. Biswas, Semi-supervised cross-modal retrieval with label prediction, IEEE Trans. Multimedia 22 (9) (2020) 2345–2353.
[4] P. Hu, H. Zhu, X. Peng, J. Lin, Semi-supervised multi-modal learning with balanced spectral decomposition, in: AAAI Conference on Artificial Intelligence, 2020, pp. 99–106.
[5] Y. Wu, S. Wang, G. Song, Q. Huang, Augmented adversarial training for cross–modal retrieval, IEEE Trans. Multimedia 23 (2021) 559–571.
[6] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach.Intell. 36 (3) (2013) 521–535.
[7] J. Wu, X. Xie, L. Nie, Z. Lin, H. Zha, Reconstruction regularized low-rank subspace learning for cross-modal retrieval, Pattern Recognit. 113 (2021) 107813.
[8] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[9] X. Ma, T. Zhang, C. Xu, Multi-level correlation adversarial hashing for cross–modal retrieval, IEEE Trans. Multimedia 22 (12) (2020) 3101–3114.
[10] S. Jin, S. Zhou, Y. Liu, C. Chen, X. Sun, H. Yao, X.-S. Hua, SSAH: semi-supervised adversarial deep hashing with self-paced hard sample generation, in: AAAI Conference on Artificial Intelligence, 2020, pp. 11157–11164.
[11] Y. Wang, B. Xue, Q. Cheng, Y. Chen, L. Zhang, Deep unified cross-modality hashing by pairwise data alignment, in: International Joint Conference on Artificial Intelligence, 2021, pp. 1129–1135.
[12] C. Sun, X. Song, F. Feng, W.X. Zhao, H. Zhang, L. Nie, Supervised hierarchical cross-modal hashing, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 725–734.
[13] G. Ding, Y. Guo, J. Zhou, Y. Gao, Large-scale cross-modality search via collective matrix factorization hashing, IEEE Trans. Image Process. 25 (11) (2016) 5427–5440.
[14] L. Wu, Y. Wang, L. Shao, Cycle-consistent deep generative hashing for cross–modal retrieval, IEEE Trans. Image Process. 28 (4) (2019) 1602–1612.
[15] W. Wang, Y. Shen, H. Zhang, Y. Yao, L. Liu, Set and rebase: determining the semantic graph connectivity for unsupervised cross-modal hashing, in: International Joint Conference on Artificial Intelligence, 2020, pp. 853–859.
[16] H. Hu, L. Xie, R. Hong, Q. Tian, Creating something from nothing: unsupervised knowledge distillation for cross-modal hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3123–3132.
[17] L. Wang, J. Yang, M. Zareapoor, Z. Zheng, Cluster-wise unsupervised hashing for cross-modal similarity search, Pattern Recognit. 111 (2021) 107732.
[18] P.-F. Zhang, Y. Li, Z. Huang, X.-S. Xu, Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval, IEEE Trans. Multimedia (2021) inpress.
[19] J. Yu, H. Zhou, Y. Zhan, D. Tao, Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing, in: AAAI Conference on Artificial Intelligence, 2021, pp. 4626–4634.
[20] Y. Zhang, W. Zhou, M. Wang, Q. Tian, H. Li, Deep relation embedding for cross–modal retrieval, IEEE Trans. Image Process. 30 (2020) 617–627.
[21] X. Wang, W. Zhu, C. Liu, Semi-supervised deep quantization for cross-modal search, in: ACM International Conference on Multimedia, 2019, pp. 1730–1739.

[22] X. Liu, G. Yu, C. Domeniconi, J. Wang, Y. Ren, M. Guo, Ranking-based deep cross-modal hashing, in: AAAI Conference on Artificial Intelligence, 2019, pp. 4400–4407.
[23] J. Zhang, Y. Peng, M. Yuan, SCH-GAN: semi-supervised cross-modal hashing by generative adversarial network, IEEE Trans. Cybern. 50 (2) (2020) 489–502.
[24] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).
[25] J. Duan, Y. Luo, Z. Wang, Z. Huang, Semi-supervised cross-modal hashing with graph convolutional networks, in: Australasian Database Conference, 2020, pp. 93–104.
[26] Z. Shen, D. Zhai, X. Liu, J. Jiang, Semi-supervised graph convolutional hashing network for large-scale cross-modal retrieval, in: IEEE International Conference on Image Processing, 2020, pp. 2366–2370.
[27] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM International Conference on Multimedia, 2010, pp. 251–260.
[28] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: ACM International Conference on Multimedia, 2014, pp. 7–16.
[29] Y. Cao, M. Long, J. Wang, Q. Yang, P.S. Yu, Deep visual-semantic hashing for cross-modal retrieval, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1445–1454.
[30] V. Erin Liong, J. Lu, Y.-P. Tan, J. Zhou, Cross-modal deep variational hashing, in: IEEE International Conference on Computer Vision, 2017, pp. 4077–4085.
[31] H.T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, R. Hong, Exploiting subspace relation in semantic labels for cross-modal hashing, IEEE Trans. Knowl. Data Eng. 33 (10) (2021) 3351–3365.
[32] S. Su, Z. Zhong, C. Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 3027–3035.
[33] D. Zhang, X.-J. Wu, Robust and discrete matrix factorization hashing for cross–modal retrieval, Pattern Recognit. 122 (2022) 108343.
[34] J. Zhang, Y. Peng, M. Yuan, Unsupervised generative adversarial cross-modal hashing, in: AAAI Conference on Artificial Intelligence, 2018, pp. 539–546.
[35] C. Li, C. Deng, L. Wang, D. Xie, X. Liu, Coupled cyclegan: unsupervised hashing network for cross-modal retrieval, in: AAAI Conference on Artificial Intelligence, 2019, pp. 176–183.
[36] X. Wang, X. Liu, Z. Hu, N. Wang, W. Fan, J.-X. Du, Semi-supervised semantic-preserving hashing for efficient cross-modal retrieval, in: IEEE International Conference on Multimedia and Expo, 2019, pp. 1006–1011.
[37] X. Liu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, M. Guo, Weakly-supervised cross-modal hashing, IEEE Trans. Big Data (2019) inpress.
[38] Q. Li, Z. Han, X.-M. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: AAAI Conference on Artificial Intelligence, 2018, pp. 3538–3545.
[39] B. Jiang, Z. Zhang, D. Lin, J. Tang, B. Luo, Semi-supervised learning with graph learning-convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11313–11320.
[40] B. Hui, P. Zhu, Q. Hu, Collaborative graph convolutional networks: Unsupervised learning meets semi-supervised learning, in: AAAI Conference on Artificial Intelligence, 2020, pp. 4215–4222.
[41] R. Xu, C. Li, J. Yan, C. Deng, X. Liu, Graph convolutional network hashing for cross-modal retrieval, in: International Joint Conference on Artificial Intelligence, 2019, pp. 982–988.
[42] W. Wang, Y. Huang, Y. Wang, L. Wang, Generalized autoencoder: a neural network framework for dimensionality reduction, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 490–497.
[43] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: ACM International Conference on Multimedia, 2017, pp. 154–162.
[44] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from national university of singapore, in: ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.
[45] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, R. Wang, Modality-specific and shared generative adversarial network for cross-modal retrieval, Pattern Recognit. 104 (2020) 107335.

Fei Wu received the PhD degree in information and communication engineering from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2016. He is currently an associate professor with the College of Automation and Artificial Intelligence, NJUPT. He has authored over 40 scientific papers, such as TPAMI, TIP, PR, CPVR, AAAI and IJCAI. His current research interests include pattern recognition and artificial intelligence.

Shuaishuai Li is currently a Master candidate in control engineering with NJUPT, Nanjing, China. His current research interests include pattern recognition and data mining.

Guangwei Gao received the PhD degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2014. Now, he is an associate professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on pattern recognition and computer vision.

Yimu Ji received the PhD degree in computer science from NJUPT, Nanjing, China, in 2006. He is a professor with NJUPT. His current research interests include intelligent driving, computer vision and big data processing.

Xiao-Yuan Jing received the Doctoral degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 1998. He is currently a professor with the School of Computer, Wuhan University, Wuhan, China. He has published more than 100 papers, such as TPAMI, TIP, TIFS, TCSVT, TMM, PR, CPVR, AAAI, IJCAI, and ICSE. His current research interests include pattern recognition and artificial intelligence.

Zhiguo Wan received the PhD degree from the School of Computing, National University of Singapore, Singapore, in 2007. He was an associate professor with the School of Computer Science and Technology, Shandong University. From 2008 to 2014, he was an assistant professor with the School of Software, Tsinghua University. He was a post-doctoral researcher with the Katholieke University of Leuven, Belgium, from 2006 to 2008. He is currently a principal investigator with Zhejiang Laboratory, Hangzhou, Zhejiang, China. His research interest includes intelligent computing.