Graph Convolutional Network Hashing for Cross-Modal Retrieval

Ruiqing Xu[1] , **Chao Li**[1] , **Junchi Yan**[2] , **Cheng Deng**[1*] and **Xianglong Liu**[3]
[1]School of Electronic Engineering, Xidian University
[2]Dept. of CSE & MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[3]Beihang University
{rqxu, li_chao}@stu.xidian.edu.cn, yanjunchi@sjtu.edu.cn, chdeng.xd@gmail.com,
xlliu@nlsde.buaa.edu.cn

Abstract

Deep network based cross-modal retrieval has recently made significant progress. However, bridging modality gap to further enhance the retrieval accuracy still remains a crucial bottleneck. In this paper, we propose a Graph Convolutional Hashing (**GCH**) approach, which learns modality-unified binary codes via an affinity graph. An end-to-end deep architecture is constructed with three main components: a semantic encoder module, two feature encoding networks, and a graph convolutional network (GCN). We design a semantic encoder as a teacher module to guide the feature encoding process, *a.k.a.* student module, for semantic information exploiting. Furthermore, GCN is utilized to explore the inherent similarity structure among data points, which will help to generate discriminative hash codes. Extensive experiments on three benchmark datasets demonstrate that the proposed GCH outperforms the state-of-the-art methods.

1 Introduction

With the development of social network and the Internet, visual data (e.g., photos and videos) have experienced an explosive growth, thus introducing renew enthusiasm in nearest neighbor search field. Cross-modal retrieval is one of the most popular approaches, which aims to search semantically similar data points in one modality (e.g., image) by using a query from another modality (e.g., text). Besides, hashing method has received wide attention because of its low storage requirements and fast query speed, which maps high dimensional multi-modal data points into common Hamming space, endowing similar cross-modal data points with similar hash codes. However, the modality gap caused by the heterogenous nature of data from different modalities still remains a challenge for accurate cross-modal hashing.

Many cross-modal hashing methods have been proposed [Bronstein *et al.*, 2010; Ding *et al.*, 2014; Shen *et al.*, 2015; Deng *et al.*, 2016], most of which adopt shallow structures and hand-crafted features. However, there is a main drawback indwelling in these methods, that the hand-crafted feature extraction procedure is independent of hash code learning procedure, limiting the capacity of achieving accurate hash codes.

On the other hand, deep network-based methods [Yang *et al.*, 2017; Deng *et al.*, 2018; Yang *et al.*, 2018; Li *et al.*, 2018; Li *et al.*, 2019] extract features with convolution neural networks and then learn hash code simultaneously in an end-to-end training fashion, which generates more accurate hash codes. Unfortunately, most of these methods design different networks for image and text data, such as two-stream network utilized in Deep Cross-Modal Hashing [Jiang and Li, 2017], and directly employ similarity matrix as semantic constraints to generate hash codes. Such naive approaches can not fully exploit semantic correlations to guide the hash code learning procedure. Furthermore, semantic structural similarities between data points can be very helpful when generating semantic-preserving hash codes, which is often ignored. Therefore, how to incorporate semantic relevance and structural similarities between different data points into hash code learning procedure is of unprecedented importance.

Data points being independent of each other is a core assumption of existing machine learning algorithms. However, this assumption does not hold for graph data where each data point (node) is related to others (neighbors) via some complex link information, which can capture the interdependence among data points. The same intuition also exists in cross-modal retrieval, since every data pair in both modalities is linked with neighboring pairs, and employing such interdependence can be beneficial to accurate retrieval. Deep structure operates on graph-structured data, such as Graph Convolutional Networks (GCNs), has attracted increasing attention because of its fine capacity of exploiting relationships between nodes [Huang and Chen, 2017; Yang *et al.*, 2019]. One of the first research on GCNs is presented in [Bruna *et al.*, 2013], many variants and improvements have been proposed since then, showing promising results in applications such as node classification [Duvenaud *et al.*, 2015; Kipf and Welling, 2016]. The basic idea of GCNs is updating one node's feature based on neighboring nodes' according to the adjacent matrix of this graph, therefore, GCN can pay attention to the semantic structure of data points via adjacency relationships, incorporating such networks in cross-modal hashing can be favorable for learning structural

*Contact Author

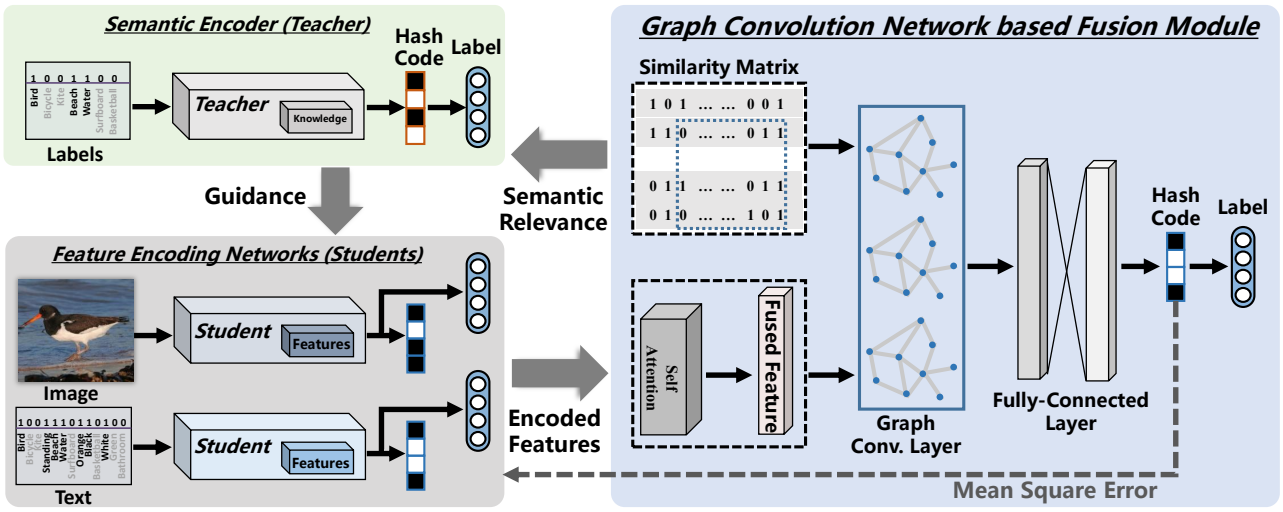

Figure 1: The proposed GCH: green, gray and blue part indicate semantic encoder, feature encoding networks and fusion module, respectively.

similarity-preserving hash codes and further boosting cross-modal retrieval performance.

In this paper, we propose a graph convolution network hashing (GCH) for cross-modal retrieval, which consists of three main components: a semantic encoder, two feature encoding networks and a graph convolution network (GCN) based fusion module. The proposed method utilizes GCN to produce an exhausted relation description between different data points, employs semantic encoder to guide the feature encoding process and preserves semantic together with structure similarities in feature learning at the same time, which is beneficial to generating more discriminative hash codes. The highlights of our work are as follows:

- We propose a novel graph convolution networks based cross-modal hashing method to alleviate the modality gap and improve cross-modal retrieval.
- To fully and effectively explore the semantic information, we train the semantic encoder to discover semantic relevance, which acts as 'teacher module' guiding feature encoding networks to learn discriminative and semantic-rich features. Then GCN is used to further enrich features with semantic structure, obtaining a beacon feature for further updating encoded features.
- Experiments on three benchmarks demonstrate that our proposed GCH notably outperforms the current state-of-the-art cross-modal hashing methods, including both traditional and deep learning based methods.

2 Related Work

Cross-modal hashing methods can be categorized into two different settings: unsupervised and supervised approaches. As an unsupervised method, in Collective Matrix Factorization Hashing (CMFH) [Ding *et al.*, 2014], unified hash codes for multiple modalities are generated using collective matrix factorization from different views.In contrast, CMSSH [Bronstein *et al.*, 2010], which is a supervised approach, presents a cross-modal hashing by preserving intra-class similarity via eigen-decomposition and boosting.

Models based on deep networks [Cao *et al.*, 2016; Li *et al.*, 2018; Jiang and Li, 2017; Cao *et al.*, 2018] are widely regarded and can better access to more discriminative features than those utilizing hand-crafted features, which leads to a boost in the performance of deep cross-modal retrieval. In recently proposed Cross-Modal Hamming Hashing (CMHH) [Cao *et al.*, 2018], favorable hash codes are generated for accurate retrieval by jointly optimizing a novel exponential focal loss and an exponential quantization loss in a Bayesian learning framework. In addition, similar to our approach, Deep Cross-Modal Hashing (DCMH) [Jiang and Li, 2017] and Self-Supervised Adversarial Hashing (SSAH) [Li *et al.*, 2018] both learn hash codes by preserving label similarity correlation, making full use of semantic information. These two methods achieve satisfying results, yet DCMH forces hash codes to keep semantic relevance for similar data points using similarity matrix in an obvious way, without paying too much attention to latent structure of cross-modal data. On the other hand, SSAH notices semantic structure lying beneath data points, but the algorithm is very time-consuming. Therefore, how to efficiently connect different modalities and explore modality consistency under the supervised information to produce favorable hash codes is the key to improve search accuracy for cross-modal hashing.

Different from existing deep network based cross-modal hashing methods, our work employs GCN to excavate data structural information and utilizes semantic encoder to extract semantic information from different modalities, which are transferred into encoded features, thus both semantic and structure similarities are well preserved, leading to optimal hash codes and better retrieval performance.

3 The Proposed GCH Model

In this work, we focus on retrieval between image and text modalities, which are two most commonly used modalities in

daily life. Figure 1 shows the flowchart of the proposed GCH model. It consists of three main components: a semantic encoder, two feature encoding networks for both modalities, and a graph convolution network based fusion module, which will all be introduced concretely in the framework subsection.

3.1 Problem Formulation

Given a cross-modal dataset $O = \{o_i\}_{i=1}^n$ with n data points, $o_i = (x_i, y_i, l_i)$, where x_i and y_i are original image and text of the i-th data point, and $l_i = [l_{i1}, \ldots, l_{ic}]$ is the ground-truth label assigned to o_i, where c is the class number. If o_i belongs to the j-th class, $l_{ij} = 1$, otherwise $l_{ij} = 0$. Specifically, we utilize multi-label similarity matrix S to denote similarity of two data points m, n: if they belong to at least one same class, $S_{mn} = 1$, otherwise $S_{mn} = 0$. The goals of cross-modal hashing are to learn a unified K-bit hash code for both modalities: $B \in \{-1, 1\}^K$, and to simultaneously preserve the original similarity between data pairs. Furthermore, to measure the similarity between two hash codes h_i and h_j, we can calculate their Hamming distance with their inner product $\langle h_i, h_j \rangle$. In order to obtain hash code for either modality, we can simply perform a non-linear transformation on the encoded feature, which can be depicted as:

$$H^* = h(f^*), \tag{1}$$

where $* = \{x, y, l\}$, $h(\cdot)$ denotes the non-linear transformation, and the obtained hash code is denoted as H^*.

3.2 Framework

In order to discover abundant semantic information in labels and transfer such information to encoded features, inspired by the idea of 'Teacher-Student' strategy, we construct a novel semantic encoder as a teacher module to fully exploit the knowledge of semantic information lying in labels, and guide the feature encoding process with these knowledge. The encoder can be formulated as follows:

$$g^l = G^l(l, \theta^l). \tag{2}$$

where θ^l is network parameter. Concretely, it is an end-to-end full-connected deep neural network. We hope the semantic encoder G^l can well preserve the similarities between semantic features and corresponding hash codes, to this end, the objective function of semantic encoder is formulated as follows:

$$\begin{aligned} \min_{\theta^l} \mathcal{L}^l &= -\alpha \sum_{i,j=1}^{n} \left(S_{ij} \Gamma_{ij}^l - \log\left(1 + e^{\Gamma_{ij}^l}\right)\right) \\ &\quad + \beta \left\| \hat{L}^l - L \right\|_F^2, \end{aligned} \tag{3}$$

where $\Gamma_{ij}^l = \frac{1}{2}(H_i^l)(H_j^l)^\top$, α and β are hyper-parameters, $\|\cdot\|_F$ is the Frobenius norm. H_{*i}^l is the predicted hash code transformed from the feature g^l and $\hat{L}^l$ is the predicted label, which can also be obtained from the feature. In Eq. (3), the first term is negative log-likelihood function, which is used to preserve similarities between features, and the second one is classification loss between the original label L and the predicted label $\hat{L}^l$. The output of semantic encoder is very helpful in guiding feature encoding networks to learn a semantic-rich feature, which favors the generation of hash codes for both modalities.

Furthermore, in order to build correlation between different modalities and further learn reliable hash codes, we construct two feature encoding networks to encode cross-modal data into common representation under the supervision of semantic encoder. For the i-th data point o_i, we model the image feature encoding function $E^x(x, \theta^x)$ with convolutional neural network to extract high-level image feature f^x. Furthermore, we construct text feature encoding network $E^y(y, \theta^y)$ with four fully-connected layers, where θ^x and θ^y are network parameters. The feature encoding networks can be written as:

$$f^* = E^*(*, \theta^*), * = \{x, y\}. \tag{4}$$

In addition, we wish to preserve the knowledge distilled from labels, *i.e.*, the semantic relevance, in encoding networks. Therefore, an end-to-end training procedure under the guidance of semantic encoder must be adopted for both encoding processes. In this way, knowledge, *i.e.*, the semantic relevance, is transferred from 'Teacher' (semantic encoder) to 'Students' (feature encoding networks). In order to introduce the supervision of semantic encoder, similar to Eq. (3), the objective function of feature encoder takes following form:

$$\begin{aligned} \min_{\theta^*} \mathcal{L}^* &= \alpha \mathcal{J}_1 + \beta \mathcal{J}_2 + \gamma \mathcal{J}_3 \\ &= -\alpha \sum_{i,j=1}^{n} \left(S_{ij} \Gamma_{ij}^{l*} - \log\left(1 + e^{\Gamma_{ij}^{l*}}\right)\right) \\ &\quad + \beta \left\| H^b - H^* \right\|_F^2 + \gamma \left\| \hat{L}^* - L \right\|_F^2. \end{aligned} \tag{5}$$

Like those defined in Eq. (3), $\Gamma_{ij}^{l*} = \frac{1}{2}(H_i^l)(H_j^*)^\top$, α, β and γ are hyper-parameters, $\|\cdot\|_F$ is the Frobenius norm. H^* are predicted hash codes for both modalities and $\hat{L}^*$ are predicted labels. Specifically, H^b is the beacon hash code generated by GCN, which contributes to encoding more discriminative features and will be discussed in the later section. It is worth noting that in Eq. (3), we encode features for each modality under the guidance of semantic encoder, which takes form of H^l in Γ_{ij}^{l*}. By doing so, semantic relevance obtained from semantic encoder is well preserved in encoded features for both modalities. Similarly to Eq. (3), here $\mathcal{J}_3$ is employed to keep categorical information by reducing the difference between $\hat{L}$ and the original label L.

To further enhance encoded features, we first need to fuse encoded features without losing too much semantic relevance. Inspired by [Vaswani *et al.*, 2017], self-attention mechanism is chosen for semantic-preserving fusing method in our work. Concretely, features from two modalities are re-weighted using features from relative modality, which can be formulated as follows:

$$f_r = \frac{1}{2}(f_{s-a}^x + f_{s-a}^y), f_{s-a}^x = f^x \times \widetilde{W}, f_{s-a}^y = f^y \times \widetilde{W}. \tag{6}$$

where $\widetilde{W} = \texttt{norm}(f^x \times f^{y\top})$ serves as the normalized weight matrix, f^x and f^y are original features from respective modality, f_{s-a}^x and f_{s-a}^y are the processed features and f_r is the fused feature. As for operative symbols, $\times$ denotes matrix inner product, and `norm` is matrix normalization.

Based in this fusion module, we can formulate out graph convolution network as follow: given N_b pairs of training

Algorithm 1 Semantic encoder guided learning for graph convolutional network hashing (GCH).

Input: Image set X; text set Y; label set L;
Output: Learned network parameters $\theta^{x,y,l,G}$;
Initialization: network parameters $\theta^{x,y,l,G}$; hyper-parameters: α, β, γ; learning rate: μ; mini-batch size: $N_b^{x,y,l} = 128$; maximum iteration number: T_{max}, iter=0;
while iter $< T_{max}$ **do**
    Update θ^l by BP algorithm.
    Update $\theta^{x,y}$ under guidance of semantic encoder.
    Update θ^G under guidance of semantic encoder, calculate beacon feature.
    Re-update $\theta^{x,y}$ using beacon feature.
end while

data points $\{x_i, y_i, l_i\}_{i=i}^{N_b}$, after feeding corresponding data points to each feature encoding networks, we will have two feature matrices $f_x \in \mathbb{R}^{N_b \times d}$ and $f_y \in \mathbb{R}^{N_b \times d}$, which will be fused using afore mentioned self-attention fusing mechanism. After obtaining the fuse semantic-rich feature f_r, it needs to be further enhanced with structural similarities to make up for the deficiency of the modality gap. To this end, we employ a multi-layer GCN, in which way in-batch features with strong latent structural relations will interact during parameter updating, leading to optimal hash codes to unify both modalities and to improve retrieval accuracy eventually.

The fused feature is fed into multi-layer GCN along with the adjacency matrix $A \in \mathbb{R}^{N_b \times N_b}$ concerning the adjacent relationships of N_b in-batch data pairs. Suggested by [Kipf and Welling, 2016], the layer-wise propagation rule of multi-layer GCN can take the following form:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}). \tag{7}$$

Here, $\widetilde{A} = A + I_N$ is the normalized adjacency matrix of the undirected graph G. Entries of A can be computed as $A(i,j) = l_i \times l_j$, here l_i is the ground truth label of data point i. I_N is the identity matrix, indicating every node is connected to itself, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ is the degree matrix of $\widetilde{A}$ and $\sigma(\cdot)$ denotes an activation function, such as `ReLU`. $H^{(l)}$ and $H^{(l+1)}$ are the input and output feature matrices of the l-th layer, representing features learned by the preceding layer and the present layer. $W^{(l)}$ acts like convolutional filter in conventional CNNs in l-th layer, whose parameters will be updated during training.

Updating parameters in GCN during training is beneficial to dynamically updating nodes' features. It can be learned from Eq. (7) that, for a node V_i, graph convolution concatenates the features of V_i's neighboring nodes by a weighted summation and simultaneously assigns new features to V_i based on $\widetilde{A}$, indicating that features of neighboring nodes are encouraged to be closer. In this way, the fused feature of two modalities are tightened by structural similarities. Accordingly, the overall objective is defined as:

$$\begin{aligned}\min_{\theta^G} \mathcal{L}^{GCN} = &-\alpha \sum_{i,j=1}^{n} \left(S_{ij}\Gamma_{ij}^G - \log\left(1 + e^{\Gamma_{ij}^G}\right)\right) \\ &+ \beta \left\| \hat{L}^G - L \right\|_F^2,\end{aligned} \tag{8}$$

where $\Gamma_{ij}^G = \frac{1}{2}(H_i^b)(H_j^b)^\top$. The output of GCN is denoted as *beacon feature*, which denotes as H^b in Eq. (8) and Eq. (5), serves as a 'beacon' in the common feature space for encoded features to approach. In this way, the structural similarities are well preserved in encoded features. The remaining parameters in Eq. (8) are as same as those in Eq. (3).

3.3 Training Strategy

Combining the four aforementioned objective functions, the overall objective function of GCH can be formulated as:

$$\mathcal{L}^{all} = \mathcal{L}^x + \mathcal{L}^y + \mathcal{L}^l + \mathcal{L}^{GCN}. \tag{9}$$

In particular, we regard them of equal importance as discussed earlier in the paper. The objective Eq. (9) is learned via iterative optimization. Concretely, the optimization order is $\mathcal{L}^l \Rightarrow \mathcal{L}^* \Rightarrow \mathcal{L}_G \Rightarrow \mathcal{L}^*$, where $* = \{x, y\}$. It is noted that network parameters are fixed when training other parts of the network and are learned by utilizing stochastic gradient descent (SGD) with a back-propagation (BP) algorithm, which are widely used in existing deep methods. Algorithm 1 outlines the whole learning algorithm in detail.

After the whole network is well-trained in the end-to-end fashion, hash codes for the unseen data points can be obtained directly by feeding the original feature into feature encoding networks:

$$b_q^{x,y} = sign(f^*(b_q; \theta^*)), \tag{10}$$

where $* = \{x, y\}$.

4 Experiments and Discussions

Three popular benchmark datasets in cross-modal retrieval: MIRFLICKR-25K [Huiskes and Lew, 2008], NUS-WIDE [Chua *et al.*, 2009], and Microsoft COCO [Lin *et al.*, 2014] are adopted to validate our proposed method. Our GCH is implemented with TensorFlow [Abadi *et al.*, 2016] and executed on a server with one NVIDIA TITAN Xp GPU.

4.1 Datasets

MIRFLICKR-25K contains 25,000 data points collected from Flickr. In total, 20,015 data points are selected in our experiment, using 10,000 data points for training and 2,000 for query. The remaining part is used for retrieval. The text for each point is represented as a 1,386-dimensional bag-of-words vector, and each point is annotated with at least one of the 24 unique labels.

NUS-WIDE contains about 269,648 web images with 81 ground truth concepts. After pruning the data without any label or tag information, a subset of 188,321 data points that belong to the 21 most-frequent concepts are selected, including 10,500 data points for training and 2,100 data points for query. The rest serves as retrieval set.

TASK	Method	MIRFLICKR-25K			NUS-WIDE			MS-COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
Image Query v.s. Text Database	SCM	0.556	0.559	0.557	0.436	0.432	0.429	0.385	0.387	0.384
	CMSSH	0.568	0.595	0.529	0.437	0.445	0.423	0.538	0.479	0.450
	STMH	0.588	0.618	0.650	0.482	0.500	0.500	0.469	0.556	0.565
	SePH	0.752	0.772	0.784	0.653	0.659	0.688	0.561	0.601	0.648
	DCMH	0.724	0.731	0.731	0.568	0.561	0.596	0.505	0.536	0.557
	OURS	**0.833**	**0.857**	**0.869**	**0.693**	**0.719**	**0.753**	**0.648**	**0.686**	**0.708**
Text Query v.s. Image Database	SCM	0.556	0.559	0.557	0.436	0.432	0.429	0.385	0.387	0.384
	CMSSH	0.568	0.595	0.529	0.437	0.445	0.423	0.538	0.479	0.450
	STMH	0.613	0.623	0.654	0.435	0.472	0.460	0.524	0.554	0.578
	SePH	0.689	0.697	0.709	0.578	0.575	0.568	0.572	0.620	0.650
	DCMH	0.764	0.749	0.780	0.558	0.591	0.616	0.549	0.572	0.605
	OURS	**0.892**	**0.910**	**0.907**	**0.732**	**0.766**	**0.761**	**0.745**	**0.797**	**0.830**

Table 1: MAP evaluation. The best accuracy is shown in boldface. The baselines are based on CNNF features.

MS-COCO contains about 80,000 training images and 40,000 validation images. 117,218 data points are used in our experiment, 10,000 data points for training and 5,000 for query. The rest of data points serve as retrieval set. Each text is represented as a 2,000-dimension bag-of-words vector.

4.2 Implementation Details

We adopt the first seven layers of CNNF [Chatfield *et al.*, 2014] neural network as image feature encoder, which is one variation of AlexNet [Krizhevsky *et al.*, 2012] pretrained on Imagenet dataset [Deng *et al.*, 2009]. For texts, a neural network with four fully-connected layers is constructed to extract high-level features, (*i.e.*, $T \rightarrow 512 \rightarrow 512 \rightarrow N$). N is number of nodes in output layer, which equals to the length of hash code K or total class label c for different datasets, depending on different types of the output. For the inputs of these networks, raw images are resized into $224 \times 224 \times 3$ and texts are represented with bag-of-words vectors.

Semantic encoder is built with a three-layer feed-forward network to project label to binary codes, (*i.e.*, $L \rightarrow 512 \rightarrow 512 \rightarrow N$), N equals to the length of hash code K or total class label c for different datasets, depending on different types of the output.

A two-layer GCN with each layer's output feature dimensions being $N_b \times 1024$ and $N_b \times K$ is employed, where N_b is batch size. Activations for these two layers are `ReLU` and `sigmoid` respectively.

Regarding the activate function used in GCH: `Sigmoid` activation is used to output predicted labels; `tanh` activation is used to output hash codes; and the rest of the layers are all uniformly activated by the `LReLU` function.

4.3 Metrics and Baselines

We adopt two common used protocols in cross-modal retrieval: Hamming ranking and hash lookup, where three evaluation criteria: Mean Average Precision (MAP) with MAP@R=500, precision-recall curve (PR curve) and Precision@N curve with N=1000, are utilized.

As for baselines, five cross-modal hashing state-of-the-art methods are compared, including STMH [Wang *et al.*, 2015], CMSSH [Bronstein *et al.*, 2010], SCM [Zhang and Li, 2014] and SePH [Lin *et al.*, 2015], which are all shallow structure based cross-modal retrieval hashing methods, and DCMH [Jiang and Li, 2017], which is a deep network based approach. We additionally compare GCH with the recently proposed CMHH [Cao *et al.*, 2018], which is also a deep network based cross-modal retrieval method.

TASK	Method	Flickr	NUS	CoCo
I$\rightarrowtail$ T	GCH-A	0.755	0.525	0.641
	GCH	**0.833**	**0.693**	**0.648**
T$\rightarrowtail$ I	GCH-A	0.811	0.510	0.741
	GCH	**0.892**	**0.732**	**0.754**

Table 2: MAP results for ablation study.

4.4 Performance Evaluation

Table 1 reports the MAP for our GCH and peer methods on three popular datasets in cross-modal retrieval. During image data processing, deep CNNF features are extracted for all shallow structure-based methods to facilitate fair competition. Our GCH takes the raw images as input. From Table 1, we can see that: 1) Compared with the shallow baselines STMH, CMSSH, SCM and SePH, our GCH achieves an average of 38%/43%, 52%/52%, 53%/62%, 11%/29% increase in MAP evaluated on MIRFLICKR-25K dataset; 2) While compared with DCMH, it can be seen that GCH can also achieve average improvement by 17%/18%. Evaluation on NUS-WIDE and MS-COCO datasets with more complex test scenarios and vast amounts of data, which are more challenging, GCH always outperforms peer methods. The main reason is that the proposed semantic encoder can well acquire the semantic information and use it to guide the encoding process for features. In addition, GCN enhances features with semantic relevance and data structure, thus more reliable hash codes can be produced and retrieval performance is improved.

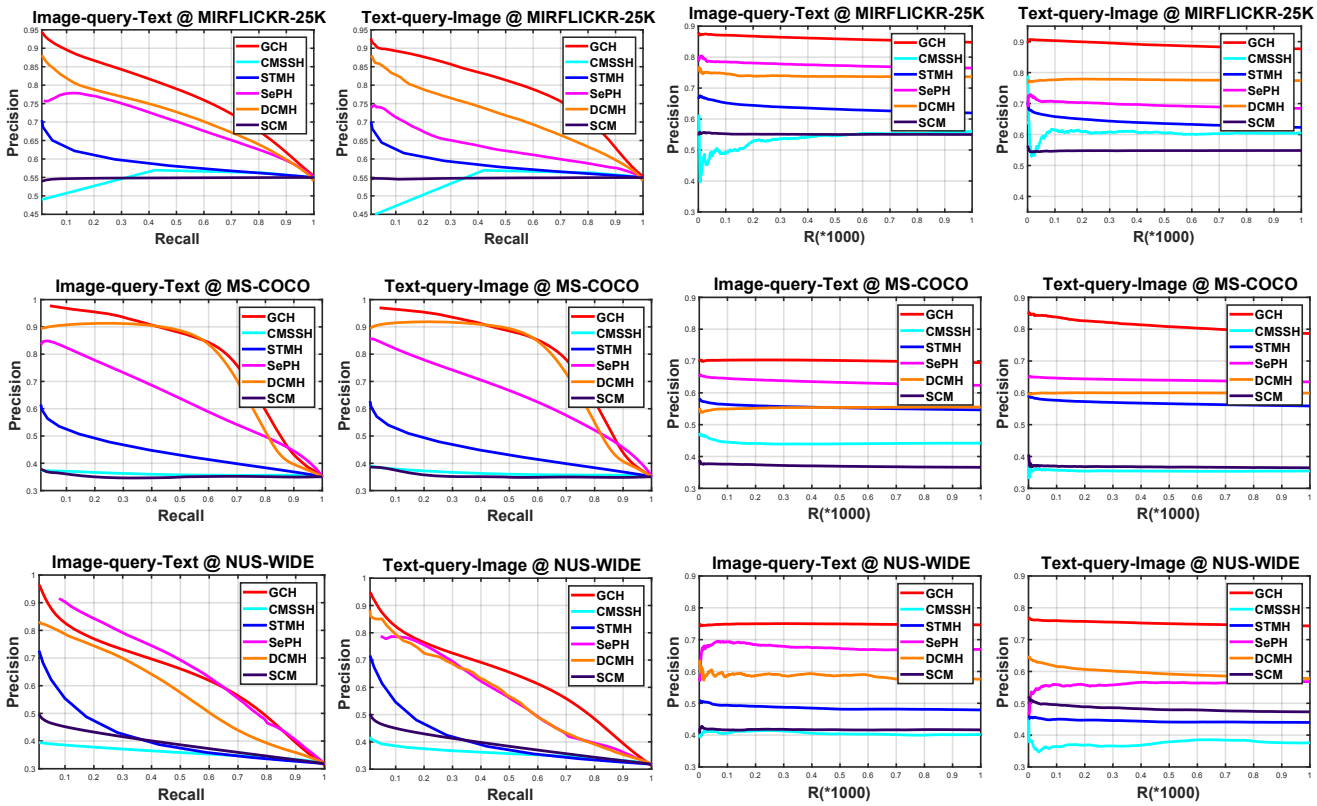

Figure 2: Precision-recall (left half) and Precision Top-1000 curves (right half). The baselines use CNN-F features. The code length is 64.

We plot the PR curve for the returned points given Hamming radius varying from 0 to 1,000 with a stepsize of 1. Left two columns in Figure 2 illustrate PR curves of all state-of-the-art methods with 64-bit hash code on three benchmark datasets, and the right two columns demonstrate Precision@1000 curves, which all show that our GCH notably outperforms all competitors.

The ablation study of GCH is quite straightforward, by simply replacing GCN with fully-connected layers, we use Table 2 to show the comparison MAP with code length 16 on three datasets. The variant of GCN is denoted as GCH-A. For simplicity, we use abbreviation for each dataset name. From the results, we can see that our method can achieve a higher performance when using the designed modules, meaning that GCN has significantly improved the overall performance.

Specially, we compare our method with Cross-Modal Hamming Hashing. CMHH is one recently proposed method on cross-modal hashing, whose source codes are not publicly available. For fair comparison, we follow the settings used in CMHH on MIRFLICKR25K dataset. Table 3 shows the experiment results, the underlined results are reported in CMHH. It can be seen that our method outperforms CMHH significantly, demonstrating the effectiveness of the proposed method.

TASK	Method	Flickr-25K
I$\rightarrowtail$ T	CMHH	<u>0.783</u>
	GCH	**0.833**
T$\rightarrowtail$ I	CMHH	<u>0.758</u>
	GCH	**0.891**

Table 3: MAP compared with CMHH.

5 Conclusion

We have presented a novel graph convolutional networks based cross-modal hashing, for large-scale cross-modal retrieval. The main contribution of our method is that we propose a GCN based hashing network for cross-modal retrieval. In addition, we utilize a novel semantic encoder to preserve rich semantic in feature encoding process. Extensive experiments on popular datasets show that our model achieves state-of-the-art performance in cross-modal retrieval task.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 61572388 and 61703327, in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2018ZDXM-GY-176, and in part by the National Key R&D Program of China under Grant 2017YFE0104100.

References

[Abadi *et al.*, 2016] Martín Abadi, Paul Barham, Jianmin Chen, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[Bronstein *et al.*, 2010] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

[Bruna *et al.*, 2013] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[Cao *et al.*, 2016] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, pages 1445–1454, 2016.

[Cao *et al.*, 2018] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *ECCV*, pages 202–218, 2018.

[Chatfield *et al.*, 2014] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, 2009.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[Deng *et al.*, 2016] Cheng Deng, Xu Tang, Junchi Yan, Wei Liu, and Xinbo Gao. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Trans. Multimed.*, 18(2):208–218, 2016.

[Deng *et al.*, 2018] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.*, 27(8):3893–3903, 2018.

[Ding *et al.*, 2014] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2083–2090, 2014.

[Duvenaud *et al.*, 2015] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*, pages 2224–2232, 2015.

[Huang and Chen, 2017] Feihu Huang and Songcan Chen. Learning dynamic conditional gaussian graphical models. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):703–716, 2017.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*. ACM, 2008.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *CVPR*, pages 3232–3240, 2017.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Li *et al.*, 2018] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, pages 4242–4251, 2018.

[Li *et al.*, 2019] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. *arXiv preprint arXiv:1903.02149*, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2015] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 2015.

[Shen *et al.*, 2015] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *CVPR*, pages 37–45, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2015] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. Semantic topic multimodal hashing for cross-media retrieval. In *IJCAI*, pages 3890–3896, 2015.

[Yang *et al.*, 2017] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 2017.

[Yang *et al.*, 2018] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. Shared predictive cross-modal deep quantization. *IEEE Trans. Neural Netw. Learn. Syst.*, (99):1–12, 2018.

[Yang *et al.*, 2019] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. *arXiv preprint arXiv:1904.13113*, 2019.

[Zhang and Li, 2014] Dongqing Zhang and Wu Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, pages 2177–2183, 2014.