IEICE TRANS. INF. & SYST., VOL.E106–D, NO.5 MAY 2023

LETTER

Modality-Fused Graph Network for Cross-Modal Retrieval

Fei WU[†a)], ***Member***, **Shuaishuai LI**[†], **Guangchuan PENG**[†], **Yongheng MA**[†], ***and*** **Xiao-Yuan JING**[††], ***Nonmembers***

SUMMARY Cross-modal hashing technology has attracted much attention for its favorable retrieval performance and low storage cost. However, for existing cross-modal hashing methods, the heterogeneity of data across modalities is still a challenge and how to fully explore and utilize the intra-modality features has not been well studied. In this paper, we propose a novel cross-modal hashing approach called Modality-fused Graph Network (MFGN). The network architecture consists of a text channel and an image channel that are used to learn modality-specific features, and a modality fusion channel that uses the graph network to learn the modality-shared representations to reduce the heterogeneity across modalities. In addition, an integration module is introduced for the image and text channels to fully explore intra-modality features. Experiments on two widely used datasets show that our approach achieves better results than the state-of-the-art cross-modal hashing methods.

key words: *cross-modal hashing, modality fusion channel, graph network*

1. Introduction

With the rapid development of the Internet, multimedia data is growing explosively. This massive multi-modal data has intricate cross-correlation relationships, and by exploiting the potential semantic associations among this multi-modal data, we can realize large-scale cross-modal data retrieval. In general, data of different modalities is heterogeneous. In recent years, to handle the modality gap issue, several cross-modal retrieval methods have been presented to explore common representations of multi-modal data.

Cross-modal hashing methods have received a lot of attention for their efficient retrieval efficiency and low storage cost. In general, existing methods can be generally categorized as unsupervised and supervised methods. Supervised information is not used in unsupervised methods [1], and the original data is projected into a common embedding space by exploring the underlying distribution and structure among multi-modal data representations. In contrast, supervised methods [2], [3] use semantic information (e.g., labels) as supervision to model the correlation between modalities to learn more discriminative hash representations, thus significantly improving retrieval performance.

In recent years, a set of deep learning based cross-modal hashing methods have been developed to make use of the tagging information for learning discriminative hashing codes. For example, Jing et al. presented the DCMH method [4], which firstly introduced deep learning into cross-modal hashing. Li et al. developed SSAH [5], by using adversarial learning to maintain semantic relevance and consistency across modalities, and designed a self-supervised semantic network that supervises the training of image and text networks. AGAH [6] uses an adversarial learning-guided multi-label attention module to enhance feature learning to learn discriminative feature representations. Xie et al. presented the consistency optimization module and the multi-task adversarial learning module in CPAH [7] for learning semantic consistency information between modalities. In DADH [8], feature alignment is performed by using adversarial training, and weighted cosine triad loss is addressed for inter-modal similarity preservation. In MLCAH [9], a multi-level correlation hashing algorithm is proposed, which encodes multi-level correlation information into hash codes by designing global and local semantic alignment mechanisms. In SAAH [10], a semantic-guided adversarial autoencoder hashing model is provided to handle inter-modal heterogeneity and improve retrieval accuracy by combining self-encoder and adversarial learning.

However, there still exists much room for improvement for existing cross-modal hashing methods. How to effectively reduce modality difference by well modeling multi-modal structure and semantic association, and simultaneously make full use of the intra-modal features has not been well studied. In this paper, inspired by the success of graph networks in various tasks, we propose a new cross-modal hashing approach called modality-fused graph network (MFGN). The contributions of our work are:

(1) The network architecture of MFGN consists of an image channel, a text channel and a modality fusion channel. The modality fusion channel fuses features of image and text channels to construct adjacency matrix, and adopts the graph network to learn the modality-shared representations. These shared representations are used to bridge cross-modal gap with the designed pairwise loss and the intra- and inter-modal discrimination loss.

(2) An integration module is designed for the image and text channels, which is used to aggregate low-level and high-level intra-modal features, such that intra-modal features are fully utilized for learning discriminative hash codes.

(3) We evaluate the effectiveness of our approach on

Manuscript received August 22, 2022.
Manuscript revised December 28, 2022.
Manuscript publicized February 9, 2023.

[†]The authors are with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China.
[††]The author is with the School of Computer, Wuhan University, Wuhan, China.
a) E-mail: wufei_8888@126.com (Corresponding author)

DOI: 10.1587/transinf.2022EDL8069

two widely used large-scale cross-modal retrieval datasets, i.e., MIRFlickr-25K [11] and NUS-WIDE [12]. The experimental results show that our approach can achieve state-of-the-art cross-modal hashing performance.

2. Our Approach

Given a multi-modal dataset $D = \{I, T\}$, where $I = [i_1, i_2, \cdots, i_N] \in \mathbb{R}^{N \times d_i}$ and $T = [t_1, t_2, \cdots, t_N] \in \mathbb{R}^{N \times d_t}$ respectively denote the feature matrices of image and text modalities, N is the total number of feature vectors of image/text modality and feature dimension $d_i \neq d_t$. Moreover, multi-label attached to the p^{th} image-text pair $o_p = (i_p, t_p)$ is represented as $l_p = \{0, 1\}^{C \times 1}$, and C denotes the number of categories. If o_p belongs to the c^{th} category, $l_{pc} = 1$, otherwise $l_{pc} = 0$. As shown in the Fig. 1, the architecture of MFGN is an end-to-end learning framework consisting of three channels, including the image channel, the text channel, and the modality fusion channel. The goal of MFGN is to learn the hash codes $B^I = \{-1, +1\}^K$ and $B^T = \{-1, +1\}^K$ for two modalities, where K is the length of the hash codes.

2.1 Modality Fusion Channel

In the modality fusion channel, Graph Convolutional Networks (GCN) [13] is used to learn modality-shared feature representations. We first concatenate image and text features to obtain $R = [I'; T']$, where I' and T' are normalizations of I and T. To fully explore multi-modal structure and semantic similarity relationship, we build an undirected graph $G = (R, Q)$, which is a graph of size N with nodes $R_i \in R$ and edges $(R_i, R_j) \in Q$. Each layer of GCN is defined as

$$H_l = \tanh(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H_{l-1} \theta_l) \tag{1}$$

where A is the adjacency matrix, $D_{ii} = \sum_j A_{ij}$, θ_l denotes the parameters to be optimized for the l^{th} layer network. In addition, H_0 is the input for GCN and $H_0 = P$. Here, $P = [I; E_T]$, where E_T is the encoded text feature of T by using a fully connected layer. E_T has the same dimension as that of I. H_2 is the output of GCN and $Z_S = H_2$.The construction of adjacency matrix A is important for graph representation learning, as it guides the network to aggregate information for each node for learning discriminative representations. Firstly, we use the Cosine distance to quantify the similarity between instances as $J_{ij} = (R_i)^T R_j$ where a larger J_{ij} indicates a greater similarity between features R_i and sample R_j, and conversely, a smaller one. However, using the Cosine distance alone is not sufficient to mine the structural information well. We thus also introduce the Euclidean distance to calculate the distance between features. The Euclidean distance based similarity matrix O is constructed as

$$O_{ij} = \exp\left(-\sqrt{\left\|R_i - R_j\right\|_2}/\rho\right) \tag{2}$$

where ρ is a scaling parameter and is set to 4 in this paper. Then, O is combined with J to obtain the total similarity metric matrix U

$$U_{ij} = (R_i)^T R_j * \exp\left(-\sqrt{\left\|R_i - R_j\right\|_2}/\rho\right) \tag{3}$$

Finally, the obtained similarity metric matrix U is combined with the label-based similarity matrix S to obtain the adjacency matrix A

$$A_{ij} = U_{ij} * S_{ij} \tag{4}$$

where $S_{ij} = 1$ means that features R_i and R_j have at least one same label and they are similar in semantics. If $S_{ij} = 0$, it means that there is no same label between R_i and R_j and they are not similar in semantics.

2.2 Integration Module

To fully explore intra-modality features, we propose an integration module to aggregate features of different dimensions to ensure the semantic integrity in the encoding process for the image and text channels. Specifically, taking the image modality as an example, the input image feature set I is encoded by fully connected layers to obtain low-dimensional features M_I, i.e., $M_I = f_I^M(I)$. We combine the low-level and high-level features to obtain the output of the module, i.e., $F_I = [I; M_I]$. In the same way, we can obtain the output $F_T = [E_T; M_T]$ for the text modality. Then, we use three fully connected layers to perform further feature mapping for image modality and text modality, i.e., $F_K^1 = f_K^1(F_K)$, $F_K^2 = f_K^2(F_K^1)$, $Z_K = f_K^3(F_K^2)$, $K \in \{I, T\}$, respectively.

2.3 Total Loss

To reduce the inter-modal heterogeneity, we define the following pairwise loss, aiming to reduce the differences between Z_I and Z_S, and between Z_T and Z_S.

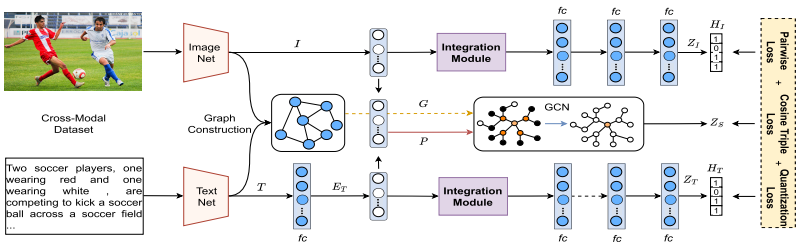

Fig. 1 The overall network architecture of our MFGN approach.

$$L_p = \|Z_I - Z_S\|_F^2 + \|Z_T - Z_S\|_F^2 \tag{5}$$

In order to maintain the similarity between hash codes, we introduce the cosine triplet loss. Each triplet consists of an anchor point and corresponding positive and negative points, denoted as $\{v, \varepsilon^+, \varepsilon^-\}$. ε^+ is a positive feature, which means this feature has at least one same label as the anchor feature. And ε^- is a negative feature, which does not have the same label as the anchor point. The cosine triplet loss is defined as follows

$$\mathcal{L}_{tri} = \sum_{i,j,k} \max\left(\cos\left(v_i, \varepsilon_k^-\right) - \cos\left(v_i, \varepsilon_j^+\right) + m, 0\right) \tag{6}$$

where m is a margin parameter. According to this definition, we define intra-modal discrimination loss L_{intra} to jointly explore discriminant information in image, text and modality fusion channels

$$\begin{aligned} L_{intra} = & \sum_{i,j,k} \max\left(\cos\left(Z_I^i, Z_I^{j-}\right) - \cos\left(Z_I^i, Z_I^{k+}\right) + m, 0\right) \\ & + \sum_{i,j,k} \max\left(\cos\left(Z_T^i, Z_T^{j-}\right) - \cos\left(Z_T^i, Z_T^{k+}\right) + m, 0\right) \\ & + \sum_{i,j,k} \max\left(\cos\left(Z_S^i, Z_S^{j-}\right) - \cos\left(Z_S^i, Z_S^{k+}\right) + m, 0\right) \end{aligned} \tag{7}$$

In addition, we also focus on cross-modal discriminant information exploration, and define the inter-modal discrimination loss L_{inter}

$$\begin{aligned} L_{inter} = & \sum_{i,j,k} \max\left(\cos\left(Z_I^i, Z_T^{j-}\right) - \cos\left(Z_I^i, Z_T^{k+}\right) + m, 0\right) \\ & + \sum_{i,j,k} \max\left(\cos\left(Z_T^i, Z_I^{j-}\right) - \cos\left(Z_T^i, Z_I^{k+}\right) + m, 0\right) \\ & + \sum_{i,j,k} \max\left(\cos\left(Z_I^i, Z_S^{j-}\right) - \cos\left(Z_I^i, Z_S^{k+}\right) + m, 0\right) \\ & + \sum_{i,j,k} \max\left(\cos\left(Z_T^i, Z_S^{j-}\right) - \cos\left(Z_T^i, Z_S^{k+}\right) + m, 0\right) \end{aligned} \tag{8}$$

To improve the retrieval efficiency and reduce the storage cost, we further map the feature representations to Hamming space to obtain the corresponding hash codes, and reduce the quantization error between the hash code and the real-valued embeddings by using the quantization loss

$$L_q = \|H_I - Z_I\|_F^2 + \|H_T - Z_T\|_F^2 \tag{9}$$

where $H_* = sign\left(Z_*\right), * \in \{I, T\}$.

Finally, the total loss function can be formulated as

$$L_{total} = L_{inter} + \alpha L_{intra} + \beta L_p + \gamma L_q \tag{10}$$

where α, β and γ are balance factors.

3. Experiments

3.1 Datasets

MIRFlickr-25K: This dataset consists of 25,000 image-text pairs, each belonging to at least one of the 24 categories. In experiments, we select the image-text pairs with at least 20 labels with a total of 20,015 pairs. The text of each image-text pair is represented as a 1,386-dimensional bag-of-words vector, while the image features are extracted using CNN-F [14] pre-trained on ImageNet with 4,096-D features.

NUS-WIDE: It is a commonly used dataset containing 269,548 image-text pairs, where each image-text pair is labeled with at least one of the 81 labels. In experiments, we select 195,834 image-text pairs with the most common 21 category labels. For each instance, the text is transformed into a 1,000-dimensional bag-of-words vector, and 4,096-dimensional features are extracted for images. The detailed division of these two datasets is shown in Table 1.

Table 1 Details of two datasets.

Dataset	Total	Train	Test	Retrieval	Labels
MIRFlickr-25K	20,015	10,000	2,000	18,015	24
NUS-WIDE	195,834	10,500	2,100	193,734	21

3.2 Implementation Details

The details of the network are as follows: the text channel uses a fully connected layer for feature encoding to obtain the feature representations with the same dimensionality as images, i.e., $d_t \rightarrow d_i$. The integration module further uses the fully connected layer to reduce the dimensionality of image/text features, i.e., $d_i \rightarrow 512$. Finally, the image/text channel learns the discriminative hash codes using three fully-connected layers, i.e, $4608 \rightarrow 1024 \rightarrow 256 \rightarrow K$. The above networks, except for the output layer, are activated using ReLu, and the output layer is activated using Tanh. The modality fusion channel uses a two-layer GCN to learn the modal-shared representations, i.e., $8192 \rightarrow 2048 \rightarrow K$.

In experiments, the hyper-parameters α, β and γ are set to 10, 10 and 0.01, respectively. In addition, the learning rate is 0.001 and we optimize the whole network using the SGD Optimizer. We focus on two cross-modal retrieval tasks: text retrieval by image query (image $\rightarrow$ text) and image retrieval by text query (text $\rightarrow$ image). We use the Mean Average Precision (MAP), i.e., the mean of the Average Precision (AP) of all queries, to evaluate the effectiveness of MFGN.

3.3 Comparison with the State-of-the-Arts

Table 2 shows the results of our approach on MIRFlickr-25K and NUS-WIDE datasets compared with state-of-the-art cross-modal hashing methods. From the table, we can see that our approach is always superior to other methods in terms of MAP for specific hash code length. Taking the hash code length of 32 as an example on the MIRFlickr-25K dataset, MFGN at least improves 0.004 = (0.812-0.808) in the case of I2T and 0.006 = (0.816-0.810) in the case of T2I on MAP, and on NUS-WIDE, MFGN at least improves 0.006 = (0.692-0.686) in the case of I2T, and at least 0.019 = (0.718-0.699) in the T2I case. The improvement is mainly due to the fact that our approach specially designs the graph

Table 2 The comparison results on the MIRFlickr-25K and NUS-WIDE datasets.

Dataset	Method	I2T			T2I			Dataset	Method	I2T			T2I		
		16 bits	32 bits	64bits	16 bits	32 bits	64bits			16 bits	32 bits	64bits	16 bits	32 bits	64bits
MIRFlickr	DCMH [4]	0.735	0.737	0.751	0.763	0.764	0.766	NUS-WIDE	DCMH [4]	0.478	0.486	0.488	0.638	0.651	0.657
	SSAH [5]	0.782	0.790	0.800	0.791	0.795	0.803		SSAH [5]	0.642	0.636	0.639	0.669	0.662	0.666
	AGAH [6]	0.770	0.795	0.805	0.763	0.773	0.797		AGAH [6]	0.652	0.655	0.657	0.631	0.645	0.640
	CPAH [7]	0.773	0.792	0.800	0.783	0.801	0.806		CPAH [7]	0.660	0.686	0.698	0.695	0.699	0.712
	DADH [8]	0.791	0.807	0.815	0.797	0.810	0.815		DADH [8]	0.636	0.667	0.672	0.657	0.674	0.707
	MLCAH [9]	0.796	0.808	0.815	0.794	0.805	0.808		MLCAH [9]	0.644	0.641	0.643	0.662	0.673	0.687
	SAAH [10]	0.792	0.796	0.815	0.795	0.803	0.806		SAAH [10]	0.628	0.646	0.656	0.651	0.663	0.659
	MFGN	**0.800**	**0.812**	**0.816**	**0.801**	**0.816**	**0.822**		MFGN	**0.665**	**0.692**	**0.708**	**0.703**	**0.718**	**0.726**

Table 3 Comparison of different variants of MFGN.

Dataset	Method	I2T			T2I		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
MIRFlickr	MFGN-f	0.789	0.804	0.815	0.787	0.803	0.812
	MFGN-i	0.790	0.808	0.814	0.800	0.811	0.818
	MFGN	**0.800**	**0.812**	**0.816**	**0.801**	**0.816**	**0.822**
NUS-WIDE	MFGN-f	0.648	0.674	0.703	0.656	0.695	0.717
	MFGN-i	0.655	0.685	0.705	0.688	0.712	0.721
	MFGN	**0.665**	**0.692**	**0.708**	**0.703**	**0.718**	**0.726**

network based modality fusion channel to deal with modality gap, and provides the integration module to effectively combine low-level and high-level intra-modal features to learn discriminative hash codes.

3.4 Discussion

In this section, we evaluate the importance of the main components in MFGH. We name the removal of the modality fusion channel as MFGN-f and the removal of the integration module as MFGN-i. Table 3 shows the results of MFGN-f, MFGN-i, and MFGN. We can be seen that both MFGN-f and MFGN-i are inferior to MFGN, which indicates the effectiveness of these two components.

This comparison implies that modality fusion channel (with the designed pairwise loss and inter-modal discrimination loss) and integration module can effectively reduce the heterogeneity between data of different modalities and improve the performance of cross-modal hashing.

4. Conclusions

In this paper, we propose a novel cross-modal hashing approach called MFGN. Aiming to uncover the commonness, the modality fusion channel learns modality-shared feature representations with graph network, which acts as an intermedium to effectively bridge the gap between image and text modalities. Furthermore, intra-modal features are fully explored by using the integration module to facilitate discriminant hash code learning. Extensive experiments on two benchmark datasets demonstrate the effectiveness of our approach. The results also demonstrate the effectiveness of the main components of MFGN.

Acknowledgments

The National Natural Science Foundation of China (No. 62076139), Open Research Project of Zhejiang Lab (No. 2021KF0AB05), 1311 Talent Program of Nanjing University of Posts and Telecommunications, and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Nos. SJCX22_0289, SJCX21_0294).

References

[1] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," IEEE Trans. Multimedia, vol.24, pp.466–479, 2021.

[2] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, and W. Wen, "Ms^{2}gah: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval," Pattern Recognition, vol.128, p.108676, 2022.

[3] X. Liu, X. Wang, and Y.-M. Cheung, "Fddh: fast discriminative discrete hashing for large-scale cross-modal retrieval," IEEE Trans. Neural Netw. Learn. Syst., vol.33, no.11, pp.6306–6320, 2022.

[4] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," IEEE Conference on Computer Vision and Pattern Recognition, pp.3232–3240, 2017.

[5] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," IEEE Conference on Computer Vision and Pattern Recognition, pp.4242–4251, 2018.

[6] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," International Conference on Multimedia Retrieval, pp.159–167, 2019.

[7] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," IEEE Trans. Image Process., vol.29, pp.3626–3637, 2020.

[8] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," International Conference on Multimedia Retrieval, pp.525–531, 2020.

[9] X. Ma, T. Zhang, and C. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," IEEE Trans. Multimedia, vol.22, no.12, pp.3101–3114, 2020.

[10] M. Li, Q. Li, Y. Ma, and D. Yang, "Semantic-guided autoencoder adversarial hashing for large-scale cross-modal retrieval," Complex & Intelligent Systems, vol.8, no.2, pp.1603–1617, 2022.

[11] M.J. Huiskes and M.S. Lew, "The mir flickr retrieval evaluation," ACM International Conference on Multimedia Information Retrieval, pp.39–43, 2008.

[12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," ACM International Conference on Image and Video Retrieval, pp.1–9, 2009.

[13] T.N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[14] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv preprint arXiv:1405.3531, 2014.