2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) | 979-8-3503-9902-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCECT57938.2023.10140500

MLSCH: Multi-layer Semantic Constraints Hashing for Unsupervised Cross-modal Retrieval

Zhaomeng Wu*
School of Cyberspace Security
Changchun University
Changchun, China
Wuzhaomeng55555@126.com

Yang Yu
Department of Facilities or Services
Shandong Haixun Information Technology Co., Ltd
Jinan, China
185517814@qq.com

Abstract—**As one of the mainstream research directions in the field of computer vision, cross-modal hash retrieval has been concerned by researchers. For real world data, unsupervised cross-modal hash retrieval is obviously more important. Aiming at the problem that some current research methods cannot convey the semantic information of high-level representation to hash code, this paper proposes a novel cross-modal hashing method, named Multi-layer Semantic Constraints Hashing for unsupervised cross-modal retrieval (MLSCH), for cross-modal retrieval. The neighbor matrix between intra-modal and inter-modal is used to guide the generation of hash codes, and the neighbor structure is applied to the feature representation of different modes, which reconstructs the cross-modal features containing structural information, and effectively improves the generation quality of hash codes. Extensive experiments show that MLSCH outperforms the current advanced cross-modal hashing methods.**

Keywords—cross-modal retrieval, unsupervised, hashing, semantic constraints

I. Introduction

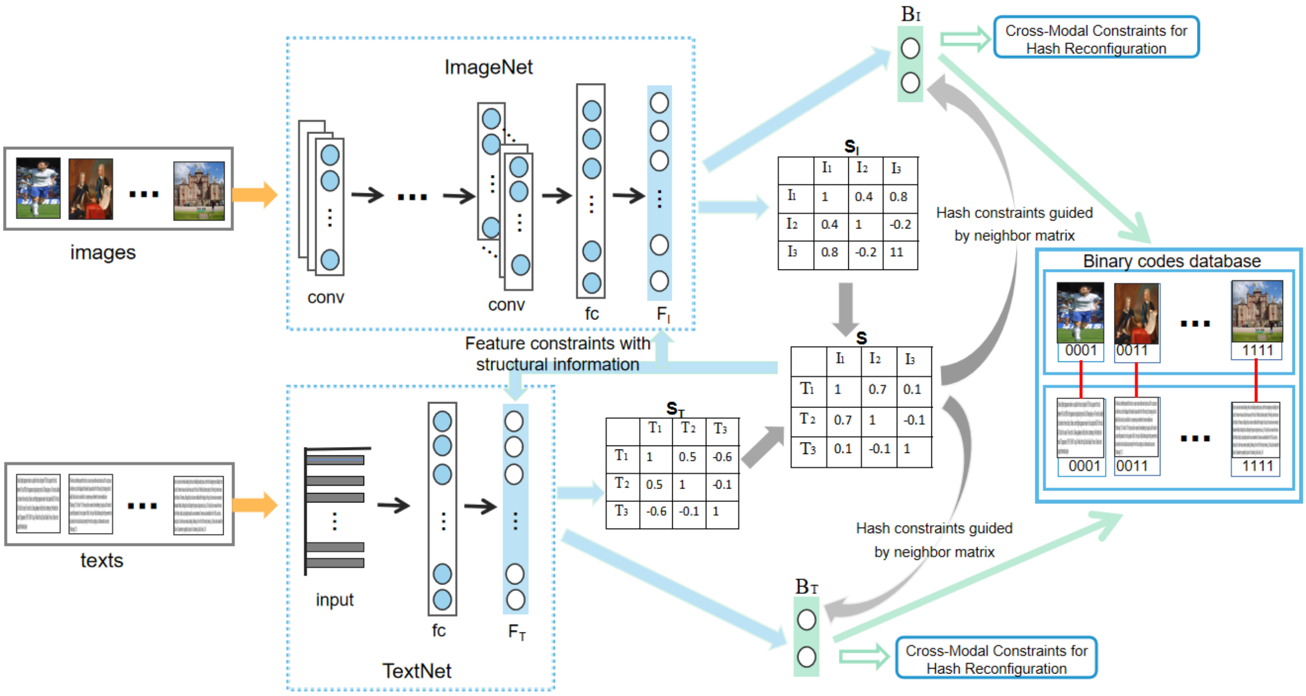
Under the background of exponential growth of multimedia data cross-modal retrieval is one of the mainstream research directions in the field of computer and information technology. The general definition of cross-modal retrieval is the use of one modality (such as text) to retrieve data from another modality (such as images). And vice versa. Of course, using images to retrieve videos, or texts to retrieve videos, etc. are all ways of cross-modal retrieval [1-4]. Cross-modal retrieval can locate and mine information more effectively, which has important application value in the web era [5,6].

It goes without saying that people are faced with huge amounts of data but want to be able to retrieve it more quickly. This demand promotes the development of cross-modal retrieval, among which the retrieval methods based on hash learning emerge one after another [7-10]. Hashing based cross-modal retrieval is mainly divided into two parts: 1) supervised cross-modal retrieval and, 2) Supervised cross-modal retrieval usually uses label information to associate image-text pairs of the same semantics and supervises binary code generation using tag information to achieve uniform encoding. Some of the representative works including Deep Cross-Modal Hashing (DCMH) [11], Self-Supervised Adversarial Hashing networks (SSAH) [12], Adversary Guided Asymmetric Hashing (AGAH) [13], Aggregation-based Graph Convolutional Hashing (AGCH) [14].

Instead of using label information, the unsupervised approach uses the spatial structure of high-level features to construct association matrices or obtains pairwise relationships between different modalities by constructing graphs and uses them as self-supervised signals to generate a unified Hamming space. Typical of these methods are Collective Matrix Factorization Hashing (CMFH) [15], Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) [16], Deep Graph-neighbor Coherence Preserving Network (DGCPN) [17].

Obviously, supervised cross-modal retrieval tends to have better performance in experiments because of the labeling information as an enrichment. But in the real world, data is so varied, and in many cases unlabeled that the time and labor cost of manual annotation would be extremely inappropriate. More recently, some unsupervised methods, such as Deep Joint-Semantics Reconstructing Hashing (DJRSH) [18], only the relation between hash codes and inter-model neighbor matrix is considered, and the correlation between hash codes and the structure of intra-modal samples is ignored. Joint-modal distribution-based similarity hashing (JDSH) [19] emphasizes ordering loss and does not take into account the importance of refactoring features.

To solve these problems, this paper proposes a new cross-modal hash method, namely, Multi-layer Semantic Constraints Hashing for unsupervised cross-modal retrieval (MLSCH). Fig.1 shows the overall framework of the proposed MLSCH. The goal of MLSCH is to make the hash codes generated by different modalities retain more semantic information and improve the performance of cross-modal retrieval. First, considering that the similarity matrix can model the relationship between instances well, MLSCH uses a distance-based neighbor matrix to construct relationships for higher-order representations of different modalities and as supervisory information to guide binary code generation, which is the first level of semantic constraints. Secondly, the algorithm acts the constructed relation matrix on the features of different modalities separately to enhance the distinguishability of the original features in order to generate distinguishable binary codes, which is the second layer of semantic constraints. Finally, the obtained binary codes are used to reconstruct the modalities such that the hash codes can learn the common representation of the same semantics among different modalities, which is the third layer of semantic constraints. Moreover, extensive experiments verify the excellent performance of our method and superior to the current advanced cross-modal hashing methods.


B_I
Cross-Modal Constraints for Hash Reconfiguration
ImageNet
images
conv
conv
fc
F_I
Hash constraints guided by neighbor matrix
Binary codes database
0001
0011
1111
0001
0011
1111
Feature constraints with structural information
texts
input
fc
F_T
TextNet
Hash constraints guided by neighbor matrix
B_T
Cross-Modal Constraints for Hash Reconfiguration

$\mathbf{S_I}$

	I1	I2	I3
I1	1	0.4	0.8
I2	0.4	1	-0.2
I3	0.8	-0.2	11

$\mathbf{S}$

	I1	I2	I3
T1	1	0.7	0.1
T2	0.7	1	-0.1
T3	0.1	-0.1	1

$\mathbf{S_T}$

	T1	T2	T3
T1	1	0.5	-0.6
T2	0.5	1	-0.1
T3	-0.6	-0.1	1

Fig. 1. The overall framework of the proposed MLSCH.

The main contributions of this paper are as follows:

- The work proposes a novel Multi-layer Semantic Constraints Hashing for unsupervised cross-modal retrieval, which can better preserve semantic information from high-level representations and generate high-quality binary codes.
- Different from the previous unsupervised cross-modal hash retrieval methods, MLSCH not only uses neighbor matrix to guide the generation of binary codes, but also apply structural information to the feature representation, and reconstruct the cross-modal features with structural information to further improve the learned common representation.
- On two commonly used cross-modal retrieval datasets, MLSCH has carried out extensive comparative experiments with other advanced methods, and has carried out erosion experiments with two of its own variants to prove the effectiveness of our method.

The rest of the structure of this paper is arranged as follows: Section II focuses on the details of our proposed MLSCH. Section III introduces our experiments on two datasets, including experimental details, parameter setting and result analysis. The conclusions are contained in section IV.

II. Proposed Method

A. Notation

In this article, bold uppercase means matrix, bold lowercase means vector. Given a multimodal training sample set containing n image-to-text pairs $\mathbf{O}=\{\mathbf{X}_i\}_{i=1}^{2}, i=\{1,2\}$. And images modalities are denoted as $\mathbf{X}_1=[\mathbf{x}_1^{(1)},\mathbf{x}_2^{(1)},\ldots,\mathbf{x}_n^{(1)}]$, texts modalities are denoted as $\mathbf{X}_2=[\mathbf{x}_1^{(2)},\mathbf{x}_2^{(2)},\ldots,\mathbf{x}_n^{(2)}]$. The deep features generated by the images are noted as $\mathbf{F_I}=\mathbf{F}_{\mathbf{X}_1}\in\mathbb{R}^{n\times d1}$, the deep features generated by the texts are noted as $\mathbf{F_I}=\mathbf{F}_{\mathbf{X}_1}\in\mathbb{R}^{n\times d2}$, where $d1,d2$ are the dimensions of the feature space of images and texts respectively. $\mathbf{B}\in\{-1,1\}^{n\times r}$ is the binary matrix to be learned, and r is the binary code length. In addition to this, method refers to the neighbor matrix of images feature space as $\mathbf{S_I}\in[-1,1]^{n\times n}$, and the matrix of texts feature space as $\mathbf{S_T}\in[-1,1]^{n\times n}$. And the semantic joint matrix $\mathbf{S}\in[-1,1]^{n\times n}$ is constructed by integrating the similarity relationship between images and texts through these two matrices, and $\tanh(\cdot)$ is a sign function, as follows,

$$\tanh(\alpha x)=\frac{e^{\alpha x}-e^{-\alpha x}}{e^{\alpha x}+e^{-\alpha x}}, x\in R \tag{1}$$

where $\alpha\in\mathbb{R}^{+}$, When α is large enough, people can clearly see that $\lim_{\alpha\to\infty}\tanh(\alpha x)=sgn(x)$. The binary hash code between $[-1,1]$ is obtained and can be optimized using back propagation.

B. Method

Deep representation of features. Multimodal data describing the same object should share the same deep semantics. Benefiting from the powerful capability of deep learning in capturing deep representations of information, the method uses deep learning techniques to map the raw features of different modalities into higher-order representations. MLSCH uses 4096-dimensional features of images extracted in pre-trained AlexNet [21], using the traditional Latent Dirichlet allocation (LDA) topic vectors or tags occurrence vectors as text features. The process of deep semantic extraction of image and text is as follows:

$$\mathbf{F_I} = f(\mathbf{I},\boldsymbol{\theta}_\mathbf{I}) \quad (2)$$
$$\mathbf{F_T} = g(\mathbf{T},\boldsymbol{\theta}_\mathbf{T})$$

Where $\mathbf{F_I} \in \mathbb{R}^{n\times d1}$, $\mathbf{F_T} \in \mathbb{R}^{n\times d2}$, $f(\cdot)$ and $g(\cdot)$ are nonlinear mapping functions for the features of images and text respectively, $\boldsymbol{\theta}_\mathbf{I}$ and $\boldsymbol{\theta}_\mathbf{T}$ are the corresponding network parameters to be learned. In this method, the images and the original semantic features from texts are passed through separate fully connected layers to obtain the same dimension of normalization, i.e., the ImageNet and the TextNet structure in Fig.1. Our work uses Eq. (2) to represent the final obtained high-level semantic feature representation.

Constructing the neighbor matrix. To maintain semantic similarity across different modalities, the method introduces similarity matrices to maintain the original nearest neighbor structure between samples. In this work, given a similarity matrix, it indicates whether two sample points are semantically related. If they are more related, the elements of the matrix are closer to 1, otherwise they are closer to -1. Therefore, the sample similarity matrix of each of the two modes can be obtained by calculating the cosine distance as follows,

$$\mathbf{S_I} = \cos(\mathbf{F_I},\mathbf{F_I^T}) = \frac{\mathbf{F_I^T F_I}}{\|\mathbf{F_I}\|\,\|\mathbf{F_I}\|} \in [-1,1]^{n\times n}$$
$$\mathbf{S_T} = \cos(\mathbf{F_T},\mathbf{F_T^T}) = \frac{\mathbf{F_T^T F_T}}{\|\mathbf{F_T}\|\,\|\mathbf{F_T}\|} \in [-1,1]^{n\times n} \quad (3)$$

Then method integrates the original neighborhood information of these two matrices to obtain the unified joint semantic matrix $\mathbf{S}$. Specifically, the method constructs a higher-order description of the relationship between two similarity matrices by weighting and summing them to obtain the new joint matrix $\mathbf{S}$,which can be stated as follows,

$$\mathbf{S} = \alpha \mathbf{S_I} + (1-\alpha)\mathbf{S_T} \quad (4)$$

where $\mathbf{S} \in [-1,1]^{n\times n}$,α is a trade-off parameter. Each element $\mathbf{S}_{\mathrm{ij}}$ in the joint matrix $\mathbf{S}$ is able to capture the semantic nearest neighbor relationship between input instances in order to obtain the structural information between samples more accurately.

Generating uniform hash codes. MLSH uses the resulting deep feature representation to directly generate the binary codes of data. For the discrete constraint problem of binary codes, the method uses a relaxation scheme to first obtain the relaxed continuous solution, and then quantize it into binary codes. To minimize quantified losses, MLSCH uses tanh function with a scale factor to produce a smoothing optimization to alleviate the gradient disappearance problem. The process is defined as follows:

$$\mathbf{B} = \tanh(\lambda \mathbf{F}) \in [-1,1]^{n\times r} \quad (5)$$

where $\lambda \in \mathbb{R}^+$, $\mathbf{F}$ is the feature representation.

Hash constraints guided by neighbor matrix. Considering the similarity preservation strategy among high-dimensional features, then the corresponding binary codes should also share the same semantic similarity. That is, in the training set, if the data samples are close to each other, the generated binary codes should also have a similar relationship. Correspondingly, if the data samples are far away from each other, then their binary codes should be extremely different. Therefore, MLSCH expects to maintain the original sample structure relationship in Hamming space, so that the binary codes of otherwise similar data remain similar and vice versa.

The first layer of semantic relations is preserved by the constructed joint similarity matrix, the method uses S as supervisory information to monitor the mutual correlation between binary codes, enabling the binary codes to preserve the semantic information of high-dimensional features:

$$\begin{aligned} Loss_1 = \mu\big(&\|\kappa\mathbf{S} - \cos(\mathbf{B_I},\mathbf{B_T})\|_F^2\big) \\ &+ \|\cos(\mathbf{B_I},\mathbf{B_T}) - \cos(\mathbf{B_T},\mathbf{B_I})\|_F^2 \\ &+ \|\kappa\mathbf{S} - \cos(\mathbf{B_I},\mathbf{B_I})\|_F^2 + \|\kappa\mathbf{S} - \cos(\mathbf{B_T},\mathbf{B_T})\|_F^2 \\ &+ \|\kappa\mathbf{S_I} - \cos(\mathbf{B_I},\mathbf{B_I})\|_F^2 + \|\kappa\mathbf{S_T} - \cos(\mathbf{B_T},\mathbf{B_T})\|_F^2 \end{aligned} \quad (6)$$

where $\cos(\mathbf{B_I},\mathbf{B_T}) \in [-1,1]^{n\times n}$ is the cosine similarity between images and texts binary matrix, and κ ,μ is a trade-off parameter.

Feature constraints with structural information. MLSCH attempts to use the rich structure in the neighbor matrix to enhance the discriminative power of the features to further generate more discriminative hash codes. The classification information of the features is continuously optimized by regressing the neighbor structure back to the feature matrix, thus maintaining the consistency of the paired data. This is indicated as follows，

$$Loss_2 = \|\mathbf{F_I} - \kappa\mathbf{S}^\mathrm{T}\mathbf{F_I}\|_F^2 + \|\mathbf{F_T} - \kappa\mathbf{S}^\mathrm{T}\mathbf{F_T}\|_F^2 \quad (7)$$

Cross-Modal Constraints for Hash Reconfiguration. For cross-modal retrieval problems, an intuitive approach is to learn the reconstructed information in order to translate the features of one modality into another modality. Usually, semantic reconstruction techniques [21,22] can effectively facilitate the matching of heterogeneous data. The method completes the cross-modal constraint of hash reconstruction by reconstructing the image binary code into the corresponding text features and the text binary code into the corresponding image features, which can be stated as follows,

$$\mathbf{F_{T\to I}} = f'(\mathbf{B_T},\rho_\mathbf{T}) \quad (8)$$
$$\mathbf{F_{I\to T}} = g'(\mathbf{B_I},\rho_\mathbf{I})$$

Where $\mathbf{F_{T\to I}} \in \mathbb{R}^{n\times d1}$, $\mathbf{F_{I\to T}} \in \mathbb{R}^{n\times d2}$, $f'(\cdot)$ and $g'(\cdot)$ are the nonlinear mapping functions required for feature reconstruction of images and feature reconstruction of texts, respectively. $\rho_{\mathbf{T}}$ and $\rho_{\mathbf{I}}$ are the corresponding network parameters to be learned. Based on the difference between the reconstructed features and the earlier features, the loss of hash reconstruction is defined as,

$$Loss_3 = \|\mathbf{F}_{\mathbf{T}\to\mathbf{I}} - \mathbf{F}_{\mathbf{I}}\|_F^2 + \|\mathbf{F}_{\mathbf{I}\to\mathbf{T}} - \mathbf{F}_{\mathbf{T}}\|_F^2 \quad (9)$$

In order to make the reconstructed features contain as much relevant structural information as possible, MLSCH will further reduce the gap between generated reconstructed features and features with structural information, as expressed below,

$$Loss_4 = \|\mathbf{F}_{\mathbf{T}\to\mathbf{I}} - \kappa\mathbf{S}^{\mathrm{T}}\mathbf{F}_{\mathbf{I}}\|_F^2 + \|\mathbf{F}_{\mathbf{I}\to\mathbf{T}} - \kappa\mathbf{S}^{\mathrm{T}}\mathbf{F}_{\mathbf{T}}\|_F^2 \quad (10)$$

After deep feature extraction, joint similarity matrix construction, binary code generation and three different layers of semantic constraints, including the similarity matrix-guided hash constraints, feature constraints with structural information, and cross-modal constraints for hash reconstruction, the final objective function to be optimized is defined as:

$$\begin{aligned} \min_{\mathbf{B_I},\mathbf{B_T}} Loss &= Loss_1 + loss_2 + loss_3 + loss_4 \\ s.t. \mathbf{B_I}, \mathbf{B_T} &\in \{-1, 1\}^{n\times r} \end{aligned} \quad (11)$$

Hash function learning. As mentioned earlier, after MLSCH obtains the binary encoding matrix containing the structural information, it needs to learn the appropriate hash function for any modality in the database. The algorithm optimizes the loss values by back propagation, and the trained parameters can be saved into the model. After that, for the query data and retrieval sets of images or texts modalities, their binary codes are computed using the following hash function,

$$\mathcal{H}_{\mathrm{I}}(x) = sgn(f(x, \boldsymbol{\theta}_{\mathbf{I}})), \mathcal{H}_{\mathrm{T}}(x) = sgn(g(x, \boldsymbol{\theta}_{\mathbf{T}})) \quad (12)$$

where $\mathcal{H}_{\mathrm{I}}(x)$a is images-oriented hash function, and$\mathcal{H}_{\mathrm{T}}(x)$ is texts-oriented function. $\boldsymbol{\theta}_{\mathbf{I}}$ and $\boldsymbol{\theta}_{\mathbf{T}}$ are the optimized parameter matrices, respectively.

III. Experiments

In this section, the proposed MLSCH is experimentally compared with several other advanced cross-modal hashing methods on two typical cross-modal datasets composed of images and texts, i.e., Wiki [23], MIRFlickr [24], and erosion experiments are carried out on Wiki dataset to illustrate the necessity of different semantic constraints in MLSCH. Specific experimental setup is described and results are analyzed.

Wiki dataset contains 2866 articles and their corresponding images in 10 categories. The dataset also holds image vectors extracted as 128-dimensional SIFT features, and 10-dimensional textual topic vectors generated by LDA model. According to the original division of the data set, 2173 image-text pairs were training set and database, and another 693 pairs were test set and query set.

MIRFlickr dataset consists of 25,000 images in 24 categories, each image is accompanied by several text labels, constituting relevant image-text pairs, and each label is annotated with at least one of the 24 category labels. Each image was described using SIFT features and each text was described using tag appearance features. 5000 image-text pairs were randomly selected as the training set, 2000 image-text pairs as the query set, and all the other data were selected as the retrieval database.

A. Compared Methods and Experimental Details

In order to confirm that our method is indeed useful, MLSCH compare the proposed approach with several advanced cross-modal methods, including DCMH [11], AGAH [13], DJRSH [18], JDSH [19], In this paper, mean average accuracy (MAP) is adopted to evaluate our method, which is a metric often used in cross-modal retrieval. This work carefully followed the experimental parameter Settings of the compared methods and verified them on two widely used cross-modal retrieval tasks: 1) Image-to-Text and, 2) Text-to-Image. Select parameters through the validation process and finally take κ=1.5,μ=1.5 for all datasets. For Wiki dataset, α=0.2 and learning rates are set to 0.005 for the ImageNet and 0.015 for the TextNet. For MIRFlickr dataset, α =0.9 and the learning rate is 0.005 for the ImageNet and 0.005 for the TextNet. Also, the SGD optimizer used has a momentum of 0.7.

B. Experiment Results

The experimental results are mainly introduced from two aspects, the first one is the comparison experiment between MLSCH and other methods. The second one is the erosion experiment between MLSCH and several variants of itself.

1) Retrieval Performance

TABLE I. THE MAP@50 RESULTS OF DIFFERENT ENCODING LENGTHS ON WIKI AND MIRFLICKR DATASETS

Task	Method	Wiki			MIRFlickr		
		16bits	***32bits***	***64bits***	***16bits***	***32bits***	***64bits***
Image-to-Text	DCMH	0.448	0.449	0.425	0.714	0.713	0.716
	AGAH	0.445	0.436	0.469	0.736	0.777	0.790
	DJRSH	0.367	0.417	0.423	0.810	0.843	0.862
	JDSH	0.369	0.429	0.436	0.833	0.850	0.880
	MLSCH	0.449	0.463	0.474	0.864	0.901	0.912
Text-to-Iamge	DCMH	0.618	0.628	0.626	0.754	0.748	0.760
	AGAH	0.623	0.630	0.650	0.747	0.773	0.782
	DJRSH	0.535	0.600	0.612	0.786	0.882	0.835
	JDSH	0.553	0.556	0.603	0.822	0.856	0.872
	MLSCH	0.628	0.652	0.658	0.844	0.876	0.886

The MAP@50 results of MLSCH with its comparison methods on Wiki dataset and MIRFlickr dataset are shown in Table I. It shows the performance of all methods when the binary code length is set to 16 bits, 32 bits, and 64 bits for the 1) image-to-text and 2) text-to-image tasks. It can be seen from Table I that: 1) MLSCH gets the best MAP results in different hash code Settings, which shows the effectiveness of our method. 2) MLSCH can compete with the supervised methods and has achieved excellent performance, which proved that our method can preserve more original structure information in binary code matrix. 3) The longer the length of binary codes, the higher the performance of all methods, indicating that the richness of semantic information in binary codes is proportional to the encoding length.

TABLE II. PERFORMANCE OF MLSCH UNDER DIFFERENT SETTINGS

Method	Configuration	*32bits*		*64bits*	
		$I \rightarrow T$	$T \rightarrow I$	$I \rightarrow T$	$T \rightarrow I$
MLSCH-1	Loss=loss1+loss3+loss4	0.395	0.632	0.408	0.630
MLSCH-2	Loss=loss1+loss2+loss3	0.417	0.568	0.398	0.596
MLSCH	Loss=loss1+loss2+loss3+loss4	0.463	0.652	0.474	0.658

2) Ablation Study

MAP results of the erosion experiments conducted on the Wiki dataset are presented in Table II. Experiments are carried out by reducing the feature constraint with structural information, i.e., MLSCH-1, and by reducing the cross-modal constraint of hash reconstruction, i.e., MLSCH-2, respectively. Then the corresponding top-K accuracy curve when the binary code length is 32 bits and 64 bits is further plotted in Fig. 2. It can be concluded that: 1) the feature constraint with structural information can optimize the structure of the original high-dimensional features, so that the semantic information within the same modality is more obvious, so that the semantic information of the binary encoding is stronger; 2) The cross-modal constraint of hash reconstruction further reconstructs the structural relationship between samples on top of reconstructing the high-order representation of different modalities, making the learned hash codes more discriminative; 3) MLSCH performs better than the other two variants in different bits, which shows the necessity of these two constraints and verify the excellence performance of proposed method.

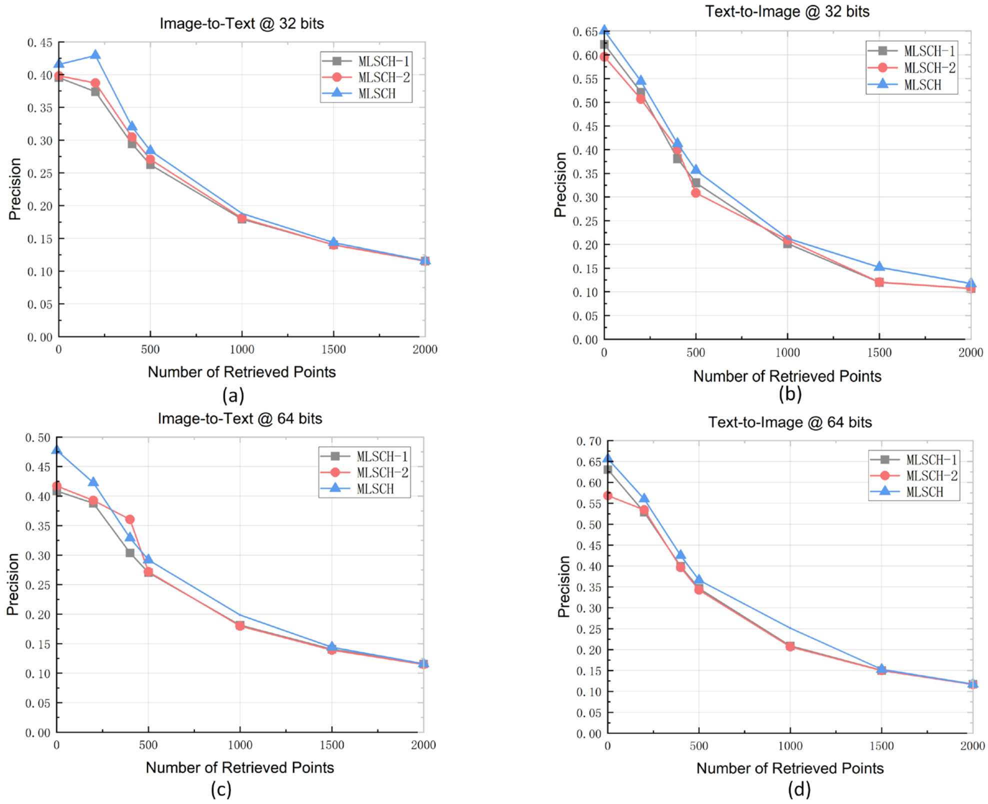


Fig. 2. Precision@top-K curve of erosion experiment on Wiki dataset

IV. Conclusion

This paper proposed a novel cross-modal hashing retrieval method, namely MLSCH. Combining the three layers of different semantic loss, our method can better utilize the structural information between samples and obtain the common representation of different modalities, and be used for learning the hash code. Our MLSCH is able to learn the relationships between modalities unsupervised, suitable for real-world applications. Extensive experiments and result analysis on two datasets show that MLSCH is superior to other advanced methods. The research plans to extend MLSCH to be able to handle unpaired data in the feature.

References

[1] Rasiwasia, Nikhil, et al. "A new approach to cross-modal multimedia retrieval." in ACM ICMR. ACM, 2010, pp. 251-260.

[2] Ngiam, Jiquan, et al. "Multimodal deep learning." in ICML-11, 2011, pp. 689-696.

[3] Xu, Ran, et al. "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework." in Proceedings of the AAAI conference on artificial intelligence, 2015, pp. 2346-2352.

[4] Elizalde, Benjamin, Shuayb Zarar, and Bhiksha Raj. "Cross modal audio search and retrieval with joint embeddings based on text and audio." in ICASSP. IEEE, 2019, pp. 4095-4099.

[5] Wang, Kaiye, et al. "A comprehensive survey on cross-modal retrieval." arXiv preprint arXiv:1607.06215, 2016.

[6] Kaur, Parminder, Husanbir Singh Pannu, and Avleen Kaur Malhi. "Comparative analysis on cross-modal information retrieval: a review." in Computer Science Review, 2021, p. 100336.

[7] Song, Jingkuan, et al. "Inter-media hashing for large-scale retrieval from heterogeneous data sources." in ACM SIGMOD. ACM, 2013, pp. 785-796.

[8] Zhang, Xi, Hanjiang Lai, and Jiashi Feng. "Attention-aware deep adversarial hashing for cross-modal retrieval." in ECCV, 2018, pp. 591-606.

[9] Xie, De, et al. "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval." in TIP, 2020, pp. 3626-3637.

[10] Yu, Jun, et al. "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing." in Proceedings of the AAAI Conference on Artificial Intelligence. 2021, pp. 4626-4634.

[11] Jiang, Qing-Yuan, and Wu-Jun Li. "Deep cross-modal hashing." in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 3232-3240.

[12] Li, Chao, et al. "Self-supervised adversarial hashing networks for cross-modal retrieval." in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 4242-4251.

[13] Gu, Wen, et al. "Adversary guided asymmetric hashing for cross-modal retrieval." in ACM ICMR. ACM, 2019, pp. 159-167.

[14] Zhang, Peng-Fei, et al. “Aggregation-Based Graph Convolutional Hashing for Unsupervised Cross-Modal Retrieval.” in IEEE T MULTIMEDIA, 2021, pp. 466–479.

[15] Ding, Guiguang, Yuchen Guo, and Jile Zhou. "Collective matrix factorization hashing for multimodal data." in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 2075-2082.

[16] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing." in Proc. 32nd AAAI Conf. Artif. Intell, 2018, pp. 1–8.

[17] Yu, Jun, et al. "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing." in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 4626-4634.

[18] Su, Shupeng, Zhisheng Zhong, and Chao Zhang. "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval." in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3027-3035.

[19] Liu, Song, et al. "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval." in ACM SIGIR. ACM, 2020, pp. 1379-1388.

[20] Krizhevsky, Alex, et al. “ImageNet Classification with Deep Convolutional Neural Networks.” in COMMUN ACM, vol. 60, no. 6, pp. 84–90. 2012

[21] Cao, Yue, et al. "Correlation autoencoder hashing for supervised cross-modal search." in ACM ICMR. ACM, 2016, pp. 197-204.

[22] Yang, Dejie, et al. "Deep semantic-alignment hashing for unsupervised cross-modal retrieval." in ACM ICMR. ACM, vol. 36, no. 3, pp. 44-52. 2020

[23] Costa Pereira, Jose, et al. “On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval.” in PAMI, 2014, pp. 521–35.

[24] Huiskes, Mark J., and Michael S. Lew. "The mir flickr retrieval evaluation." in ACM ICMR. ACM, 2008, pp. 39-43.