Critical Separation Hashing for Cross-Modal Retrieval

Zening Wang[1(✉)], Yungong Sun[1], Liang Liu[2], and Ao Li[1]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
1335028185@qq.com

[2] School of Electrical Engineering and Computer Science, Pennsylvania State University, New York, USA

Abstract. With the development of Internet technology, unimodal retrieval techniques are no longer suitable for the current environment, and mutual retrieval between multiple modalities is needed to obtain more complete information. Deep hashing has clearly become a simpler and faster method in cross-modal hashing. In recent years, unsupervised cross-modal hashing has received increasing attention. However, existing methods fail to exploit the common information across modalities, thus resulting in information wastage. In this paper, we propose a new critical separation cross-modal hashing (CSCH) for unsupervised cross-modal retrieval, which explores the similarity information across modalities by highlighting the similarity between instances to help the network learn the hash function, and we carefully design the loss function by introducing the likelihood loss commonly used in supervised learning into the loss function. Extensive experiments on two cross-modal retrieval datasets show that CSCH has better performance.

Keywords: Unsupervised · Cross-modal · Smilarity Matrix

1 Introduction

Along with the rapid development of the Internet and the popularity of smart devices and social networks, multimodal data has exploded on the Internet [1]. Multimodal data is simply a representation of the same thing in different modalities. How to follow one modality to retrieve other modalities becomes the key to searching information, which makes cross-modal retrieval emerge. General cross-modal retrieval methods use generic real values from different modalities to retrieve each other, but drawbacks such as high computational complexity and storage inefficiencies limit their use.

Cross-modal hashing methods have gained increasing interest due to the efficiency of storing binary hash codes and the simplicity of calculating Hamming distances [2], which map modal features to the same Hamming space for retrieval. In general, unsupervised methods using only information from the input image-text pair to mine out potential relationships and supervised methods using semantic labels to reduce the modal gap, which in turn aids hash code learning and obtains better performance. However, in most

A. Li et al. (Eds.): 6GN 2022, LNICST 504, pp. 171–179, 2023.
https://doi.org/10.1007/978-3-031-36011-4_15

cases, the data obtained does not have labels to work with, so unsupervised methods are more convenient [3].

The emergence of deep neural networks has facilitated the development of cross-modal hashing, and deep neural networks have a stronger semantic representation capability, which helps in the learning of further hash codes. There are still problems to be solved in unsupervised cross-modal hashing. The creation of similarity matrices requires uniform calculation of pairwise distances between different features, such as Hamming distance. In the method of constructing similarity matrices from features extracted by pre-training networks, the similarity matrix is constructed only by the direct relationship of features or by offsetting certain similar values, which is not ideal for subsequent use as a supervised matrix to learn hash codes. To solve the above problem, we propose a Critical Separation Cross-Modal Hashing method to assist in the construction of the similarity matrix, enhance or weaken the connection between features depending on the degree of similarity between instances to guide the learning of hash codes. The contributions of this paper are as follows:

- We designed a new similarity matrix, called Critical Separation Cross-Modal Hashing (CSCH), for unsupervised cross-modal retrieval. Supervised learning of hash functions using more accurate similarity matrices.
- Experimental results on two widely used retrieval datasets show that CSCH consistently outperforms other advanced techniques.

2 Related Work

2.1 Supervised Cross-Modal Hashing

Supervised cross-modal hashing methods have made some progress in cross-modal retrieval by using label information to learn how to generate hash codes. Through continuous development, benchmarks for supervised cross-modal retrieval have been improved somewhat. DCMH integrates feature learning and hash code learning into the same framework with deep neural networks [4]. THN learns cross-modal correlations jointly from auxiliary datasets, employing RNN networks and aligning the data distribution of the auxiliary dataset with the data distribution of the query or database domain to generate compact transfer hash codes for efficient cross-modal retrieval [5]. SSAH incorporates adversarial learning into cross one of the early attempts at cross-modal hashing uses adversarial networks to maximise the semantic relevance [6]. DLFH is based on a discrete latent factor model approach that allows direct learning of binary hash codes for cross-modal hashing and is suitable for cross-modal similarity search [7]. GCH learns modally uniform binary codes via affinity graphs and graph convolutional networks are used to explore inherent similarity between data points structure [8].

2.2 Unsupervised Cross-Modal Hashing

Unsupervised cross-modal hashing has no labelling information to work with and can only be learned from the data. There are two broad approaches to learning, one is shallow cross-modal hashing; the other is using deep neural networks to learn from

the data. The shallow cross-modal hashing framework, CVH extends spectral hashing to multi-modal scenarios, maintaining semantic consistency of modalities in the same space [9]. LSSH extracts the intrinsic features under different modes by sparse coding and matrix decomposition [10]. CMFH goes a step further from both by learning a unified hash code through collective matrix decomposition and latent factor model, and at the same time, multiple information sources can be integrated to improve the retrieval accuracy during retrieval [11]. The above methods are not very effective in retrieval due to the shortcomings of containing insufficient information and the time-consuming and laborious extraction process, as all the features are produced manually.

With the further development of neural networks, it compensates for the inability of shallow structures to fully exploit the non-linear relationships between different modalities. UGACH uses the unsupervised representation learning capability of GAN to exploit the underlying streaming structure to maintain similarity between data [12]. DJSRH proposes a new joint semantic affinity matrix that integrates raw neighborhoods from different modal information to capture the potential connections of the input different modal instances [13]. DSAH makes full use of co-occurring image-text pairs and designs semantic alignment loss functions to maintain consistency between the input features and the hash codes output by the network [14]. JDSH refines the joint modal similarity matrix in a weighted manner based on the original feature distribution [15]. However, the unsupervised method proposed above still suffers from the problem of inaccurate similarity, thus obtaining only sub-optimal retrieval of the Hamming space (Fig. 1).

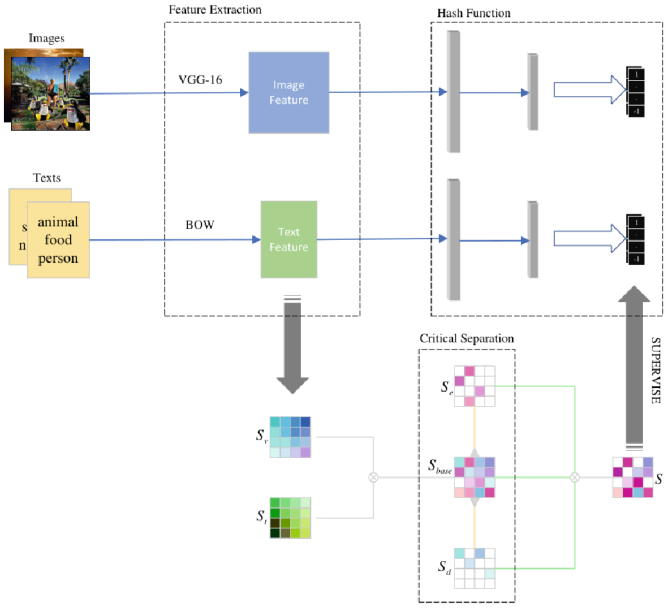

Fig. 1. The framework of CSCH.

3 The Proposed Method

3.1 Problem Definition

We assume that in the training set, $v_i \in \mathbb{R}^{d_v}$ and $t_i \in \mathbb{R}^{d_t}$ are the image and text features of the i_th instance. We use $\boldsymbol{V} = \{v_i\}_{i=1}^{N}$ represent the initial image and $\boldsymbol{T} = \{t_i\}_{i=1}^{N}$ represent the initial text features, respectively. The goal of this method is to learn the hash function $f_v(v; \theta_v)$ and $f_t(t; \theta_t)$, where θ_v, θ_t for parameters of ImgNet and TextNet. We discard Hamming similarity and use cosine values to measure the degree of similarity between instances, defined as:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}, \tag{1}$$

where $\|\cdot\|$ denotes the l_2 norm of the vector and the Frobenius norm of the matrix.

3.2 Separation of Similarity Matrix

Since there is no labelling information available, it is not possible to directly construct pairwise similarity matrices to represent the relationships between instances. We can use a pre-trained deep neural network to extract features, and a deep perceptron to extract useful semantic information to construct the similarity matrix. We use $\boldsymbol{V} = \{v_i\}_{i=1}^{N}$ to construct a similarity matrix between images $\boldsymbol{S}_v = \left\{s_{ij}^{v}\right\}_{i,j=1}^{N}$, where s_{ij}^{v} is $\cos\left(v_i, v_j\right) = \frac{v_i \cdot v_j^T}{\|v_i\| \cdot \|v_j\|}$, representing cosine similarity between image features $v_i, v_j \in \mathbb{R}^{d_v}$, and $\boldsymbol{T} = \{t_i\}_{i=1}^{N}$ to construct a cosine similarity matrix between images $\boldsymbol{S}_t = \left\{s_{ij}^{t}\right\}_{i,j=1}^{N}$, where s_{ij}^{t} is $\cos\left(t_i, t_j\right) = \frac{t_i \cdot t_j^T}{\|t_i\| \cdot \|t_j\|}$, representing cosine similarity between text features, $t_i, t_j \in \mathbb{R}^{d_t}$.

First, we weight the fusion of $\boldsymbol{S}_v$ and $\boldsymbol{S}_t$, which represent the degree of similarity between images and text, and use them as the initial similarity matrix

$$\boldsymbol{S}_{base} = \frac{1}{2}(\boldsymbol{S}_v + \boldsymbol{S}_t), \tag{2}$$

where $\boldsymbol{S}_{base} = \left\{s_{ij}\right\}_{i,j=1}^{N}$, $s_{ij} \in [0, 1]$, which maintains the cosine features of the image-text pairs, relates the relationship between the different feature pairs, we believe that the two modalities should have the same effect on the initial similarity matrix.

In the similarity critical separation section, we start by presenting several cases of values in $\boldsymbol{S}_{base}$:

$$\boldsymbol{S}_{base} \begin{cases} \textit{Strong} & s_{ij} \in [1-\sigma, 1) \\ \textit{Normal} & s_{ij} \in (0+\sigma, 1-\sigma), \\ \textit{Weak} & s_{ij} \in (0, 0+\sigma] \end{cases} \tag{3}$$

where σ is the range parameter, controling the critical range to be divided in the similarity matrix. When Normal, this part of the instances is generally similar and does not favour

either side, and we do not treat it. When it is Strong, it is called Strong similarity, this part of the instances are strongly similar to each other and can be easily distinguished at optimisation time, we choose to enhance this part to improve the accuracy at retrieval time, when it is Weak, we call it No similarity, this part of the instances are almost unrelated to each other and have little impact on optimisation, we want to further weaken the impact of this part on retrieval. The above three definitions of similarity are helpful for us to further investigate the similarity relationship between instances. In order to better generate distinguished hash codes, we need to further process the matrix $\boldsymbol{S}_{base}$ further.

First, we need to remove the extraneous parts of $\boldsymbol{S}_{base}$:

$$\boldsymbol{S}_e = \boldsymbol{S}_{base} + \boldsymbol{S}_{base}(Strong), \tag{4}$$

$$\boldsymbol{S}_d = \boldsymbol{S}_{base} - \boldsymbol{S}_{base}(Weak), \tag{5}$$

where $\boldsymbol{S}_e$ is the part we need to enhance in $\boldsymbol{S}_{base}$ and $\boldsymbol{S}_d$ is the part we need to weaken in $\boldsymbol{S}_{base}$. After normalization, we combine it with $\boldsymbol{S}_{base}$ to obtain the final similarity matrix S:

$$\boldsymbol{S} = (1-\alpha)\boldsymbol{S}_{base} + \alpha\boldsymbol{S}_e - \beta\boldsymbol{S}_d, \tag{6}$$

where α, β indicates the weighting of the parts, and in order to map $\boldsymbol{S}$ to [−1,1], we simply pass $\boldsymbol{S} = 2\boldsymbol{S} - 1$.

3.3 Objective Functions

This paper proposes comprehensive similarity preservation loss and separation loss by helping neural networks learn hash functions in three ways: distance similarity preservation, intra- and inter-modal consistency preservation, and likelihood similarity preservation.

We define, $\boldsymbol{Z}_v = f_v(v;\theta_v), \boldsymbol{Z}_t = f_t(t;\theta_t)$, $\boldsymbol{Z}_v, \boldsymbol{Z}_t$ as the real-valued eigenmatrices, then the similarity matrix for the same modality as $\boldsymbol{S}_c(\boldsymbol{Z}_v, \boldsymbol{Z}_v), \boldsymbol{S}_c(\boldsymbol{Z}_t, \boldsymbol{Z}_t)$, the similarity matrix for different modalities as $\boldsymbol{S}_c(\boldsymbol{Z}_v, \boldsymbol{Z}_t), \boldsymbol{S}_c(\boldsymbol{Z}_t, \boldsymbol{Z}_v)$, the distance similarity preservation loss as L_{dis}, and the intra- and inter-modal consistency loss as L_{in}, respectively.

$$L_{dis} = \sum_{p,q} \left\| \boldsymbol{S}_c(\boldsymbol{Z}_p, \boldsymbol{Z}_q) - \boldsymbol{S} \right\|_F, \tag{7}$$

$$L_{in} = \sum_{p,q,p_1,q_1} \left\| \boldsymbol{S}_c(\boldsymbol{Z}_p, \boldsymbol{Z}_q) - \boldsymbol{S}_c(\boldsymbol{Z}_p, \boldsymbol{Z}_q) \right\|_F, \tag{8}$$

where $p, q, p_1, q_1 \in \{v, t\}$. $\boldsymbol{S}$ is the similarity matrix computed from (6). $\|\cdot\|_F$ is the Frobenius parametrization. The likelihood similarity loss L_l is defined as:

$$L_l = -\sum_{i,j=1}^{N} \left(\boldsymbol{S}_{ij}\Theta_{ij} - \log(1 + e^{\Theta_{ij}}) \right), \tag{9}$$

where $\Theta = {}^{1}/_{2}\boldsymbol{Z}_v\boldsymbol{Z}_t^T$.

From the definition of the likelihood function, it is clear that when the value is large, two instances should have a high probability of being similar, and vice versa. Also, quantifying the similarity between instances can be turned into the problem of calculating the size of the inner product of the original features of the instances. The critical separation matrix reflects whether there is similarity between two instances and is used as supervisory information in L_{dis} instead of labels.

For maintaining intra- and inter-pattern consistency, data from different patterns of the same instance should have strong similarity, regardless of the pattern in which the data are in, as an inherent data relationship within and between patterns. Therefore, we expect intra- and inter-modal consistency to mitigate the errors arising from critical separation for some instances.

In summary, the loss function L is defined as:

$$L = L_{dis} + L_{in} + \eta L_l, \tag{10}$$

where η is the parameter used to adjust the importance of the likelihood similarity loss.

4 Experiments

4.1 Datasets

We use two public datasets. NUS-WIDE [16] and MIRFlickr-25K [17], where 10 classes commonly used in the NUS-WIDE dataset are used as our original dataset, with a total of 186,577 image text pairs. We select 2000 data pairs from them as our query set. Then we select 5000 from the remaining 166,577 data pairs as our training set. In addition, to reduce the retrieval time, we select 10,000 from the remaining data pairs as our retrieval set.

MIRFlickr-25K has 20,015 image text pairs left after removing the problematic data. We select 2000 to form the query set. The rest are used as the retrieval set, from which 5000 are selected as the training set.

For the image features, the original image is preprocessed (e.g., scaled, cropped, flipped, etc.) and fed with the trained VGG_16, keeping the features of the last fully connected layer as the input to ImageNet. The text is represented by the features of 1000-dimensional BoW and 1386-dimensional BoW, respectively.

4.2 Evaluation Metrics

We use the common mAP to measure the retrieval ability of the method. The average precision AP is defined as:

$$AP = \frac{\sum_{q=1}^{R} P(q)rel(q)}{\sum_{q=1}^{R} rel(q)}, \tag{11}$$

where $rel(q) = 1$ if the item at rank q is relevant, $rel(q) = 0$ otherwise.$P(q)$ denotes the precision of the result ranked at q. When querying, we take the average of all AP to get mAP. The experimental results are shown in Fig. 2.

Table 1. Performance comparison of ten UCMH methods on two public datasets.

		MIRFlickr-25K				NUS-WIDE			
Task	Method	8-bit	16-bit	32-bit	64-bit	8-bit	16-bit	32-bit	64-bit
I2T	CVH	0.573	0.580	0.579	0.579	0.371	0.379	0.378	0.377
	FSH	0.581	0.590	0.597	0.597	0.391	0.400	0.414	0.424
	CMFH	0.574	0.588	0.592	0.594	0.479	0.483	0.488	0.486
	LSSH	0.627	0.630	0.634	0.631	0.463	0.475	0.484	0.474
	UGACH	0.674	0.686	0.695	0.702	0.532	0.548	0.563	0.567
	DJSRH	0.646	0.666	0.678	0.699	0.496	0.513	0.535	0.566
	UKD-SS	0.693	0.700	0.706	0.709	0.554	0.564	0.557	0.561
	DSAH	0.690	0.701	0.712	0.722	0.560	0.569	0.576	0.583
	JDSH	0.651	0.669	0.683	0.698	0.548	0.554	0.561	0.582
	CSCH	**0.721**	**0.735**	**0.747**	**0.752**	**0.572**	**0.591**	**0.600**	**0.611**
T2I	CVH	0.572	0.580	0.579	0.580	0.369	0.378	0.378	0.379
	FSH	0.571	0.589	0.595	0.595	0.390	0.395	0.408	0.417
	CMFH	0.588	0.590	0.595	0.598	0.483	0.487	0.488	0.493
	LSSH	0.612	0.621	0.628	0.626	0.462	0.476	0.481	0.477
	UGACH	0.689	0.692	0.698	0.699	0.543	0.557	0.562	0.580
	DJSRH	0.672	0.683	0.694	0.717	0.532	0.546	0.561	0.583
	UKD-SS	0.694	0.704	0.705	0.714	0.580	0.587	0.583	0.583
	DSAH	0.697	0.707	0.713	0.728	0.577	0.589	0.601	0.609
	JDSH	0.682	0.686	0.699	0.716	0.567	0.572	0.586	0.605
	CSCH	**0.720**	**0.733**	**0.745**	**0.750**	**0.588**	**0.604**	**0.612**	**0.621**

4.3 Implementation Details

For image and text features, we use two fully connected layers each as sub-networks to learn the representation of hash codes. Only the relevant parameters of the sub-networks are updated during training. For critical hyperparameters, first, we set α from 0.1 to 0.9 at an increment of 0.1 per step and set β from 1 to 0.001 with 10 times increments per step. For simplicity, we change the range we set k from 2500 to 4900 at an increment of 500. Finally, we determine the values of the parameters of the negative log-likelihood function η, we set η from 1 to 0.0001 with 10 times increments per step. We use the mAP results of the query set to determine the final parameters, which are: $\alpha = 0.3$, $\beta = 0.01$, $k = 4900$, $\lambda = 4000$, $\eta = 0.001$.

4.4 Results Analysis

Table 1 shows the mAP results on our proposed CSCH and other baseline methods. We can find that the best results are obtained for CSCH with different number of bits of hash codes. Our conclusions are as follows:

(1) Among all the methods on this dataset, CSCH achieves the best results on both the image query text task(I2T) and the text query image(T2I) task. This indicates that our CSCH improves accuracy by optimising the similarity matrix.

(2) As can be seen in Table 1, CSCH obtained the most recent results on this dataset. For I2T and T2I, CSCH outperformed the previous best model (DSDH) on 8, 16, 32 and 64 bits on both datasets by 4.5%, 4.9%, 4.9%, 4.1% and 3.3%, 3.7%, 4.5%, 3.0% on MIRFlickr, respectively, and on NUS-WIDE by were 2.1%, 3.9%, 4.2%, 4.8% and 1.9%, 2.5%, 1.8%, 2.0%.

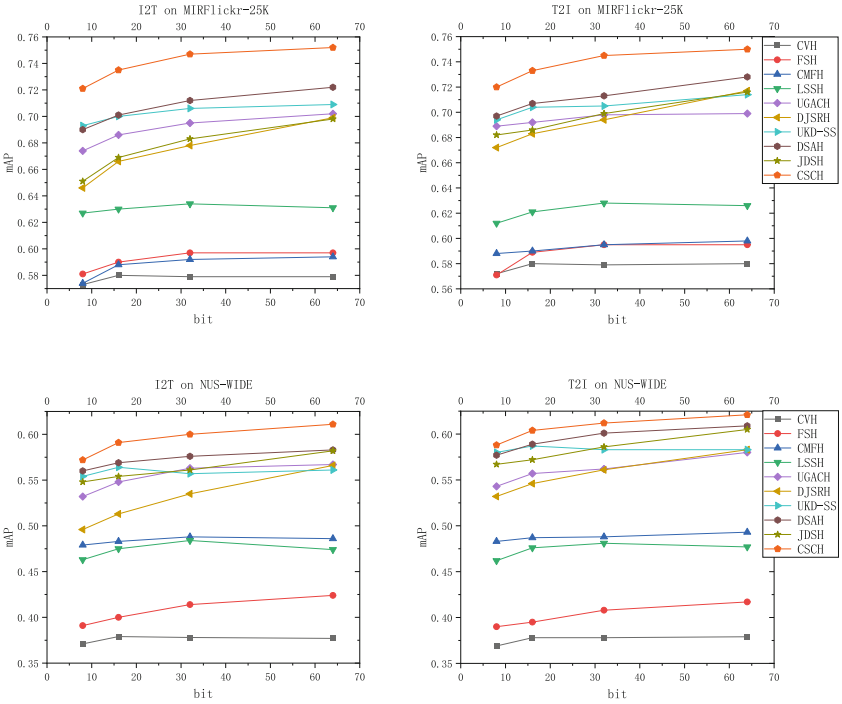

Fig. 2. The mAP results for both datasets.

5 Conclusion

In this paper, we have proposed a new unsupervised cross-modal retrieval method. Specifically, CSCH proposes 1) constructing a similarity matrix by critical separation to supervise the learning of hash functions; 2) using three complementary loss functions to aid the learning of hash functions, and specifically, invoking the likelihood loss commonly used in supervised learning to achieve stability in maintaining similarity learning. 3) A comparison with other superior techniques on two publicly available large cross-modal retrieval datasets shows that CSCH is superior and proves its superiority.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62071157, National Key Research and Development Programme

2022YFD2000500 and Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011.

References

1. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics preserving hashing for cross-view retrieval. In: CVPR, pp. 3864–3872 (2015)
2. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 415–424. ACM, 2014
3. Gu, W., Gu, X., Gu, J., Li, B., Xiong, Z., Wang, W.: Adversary guided asymmetric hashing for cross-modal retrieval. In: ICMR, pp. 159–167. ACM (2019)
4. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: IEEE CVPR, pp. 3232–3240 (2017)
5. Cao, Z., Long, M., Yang, Q.: Transitive hashing network for heterogeneous multimedia retrieval. CoRR,2016,abs/1608.04307
6. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of IEEE CVPR, pp. 4242–4251 (2018)
7. Jiang, Q., Li, W.: Discrete latent factor model for cross-modal hashing. In: IEEE Transactions on Image Processing, pp. 3490–3501 (2019)
8. Zhou, X., et al.: Graph convolutional network hashing. IEEE Trans. Cybern. 1460–1472 (2020)
9. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: Twenty-Second International Joint Conference on Artificial Intelligence (2011)
10. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval, pp. 415–424 (2014)
11. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 2075–2082 (2014)
12. Zhang, J., Peng, Y., Yuan, M.: Unsupervised generative adversarial cross-modal hashing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
13. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3027–3035 (2019)
14. Yang, D., Wu, D., Zhang, W., Zhang, H., Li, B., Wang, W.: Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 44–52 (2020)
15. Liu, S., Qian, S., Guan, Y., Zhan, J., Ying, L.: Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1379–1388 (2020)
16. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp. 1–9 (2009)
17. Huiskes, M.J., Lew, M.J.: The MIR flickr retrieval evaluation. In: Multimedia Information Retrieval. ACM, pp. 39–43 (2008)