Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval

Shupeng Su[†] Zhisheng Zhong[†] Chao Zhang[*]
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
{sushupeng, zszhong, c.zhang}@pku.edu.cn

Abstract

Cross-modal hashing encodes the multimedia data into a common binary hash space in which the correlations among the samples from different modalities can be effectively measured. Deep cross-modal hashing further improves the retrieval performance as the deep neural networks can generate more semantic relevant features and hash codes. In this paper, we study the unsupervised deep cross-modal hash coding and propose ***D****eep* ***J****oint-***S****emantics* ***R****econstructing* ***H****ashing (DJSRH), which has the following two main advantages. First, to learn binary codes that preserve the neighborhood structure of the original data, DJSRH constructs a novel joint-semantics affinity matrix which elaborately integrates the original neighborhood information from different modalities and accordingly is capable to capture the latent intrinsic semantic affinity for the input multi-modal instances. Second, DJSRH later trains the networks to generate binary codes that maximally reconstruct above joint-semantics relations via the proposed reconstructing framework, which is more competent for the batch-wise training as it reconstructs the specific similarity value unlike the common Laplacian constraint merely preserving the similarity order. Extensive experiments demonstrate the significant improvement by DJSRH in various cross-modal retrieval tasks.*

1. Introduction

Cross-modal retrieval is a classic scenario which aims to search the semantic relevant samples from different modalities, e.g., using a text description to retrieve the relevant images. Owing to the explosive increase of the multimedia data, hashing based cross-modal methods which encode the correlative samples with similar binary features are gaining importance due to the high efficiency of the binary vectors in storage and the mutual Hamming distance computation. Although the fundamental idea is applicable to any combination of content modalities, we focus on the image-text cross-modal retrieval in this paper which has been a compelling research topic in the computer vision community recently [24, 8, 18, 2, 36].

Since the heterogeneity (a.k.a. the modality gap) confines the direct measurement of the similarity among the samples from different modalities, cross-modal hashing proposes to embed the original data into a common binary hash space, in which the correlations across different modalities can be effectively and efficiently measured with their Hamming distance. Specifically, traditional cross-modal hashing can be grouped into the supervised and unsupervised categories. Unsupervised methods [16, 25, 6, 37] only utilize the co-occurrence information of the input image-text pair to maximize their correlation in the common hash space. The supervised ones [20, 1, 34, 30, 12] can further exploit the semantic labels to learn more consistent hash codes for the semantic relevant cross-modal data, which significantly mitigate the modality gap and achieve superior retrieval performance.

Deep cross-modal hashing makes a further development with the remarkable competence of the deep neural networks to generate more semantic relevant features [15, 33] which facilitates to learn more consistent hash codes subsequently. However, compared with the supervised deep cross-modal hashing which has been widely studied [3, 14, 24, 8, 18, 2, 36], the unsupervised field that our paper focuses on lacks sufficient explorations. Among the related research, UDCMH [31] is one of the latest unsupervised deep cross-modal hashing methods that integrates the graph Laplacian constraint term into the network training. It explicitly constrains the hash codes to preserve the neighborhood structure of the original data and consequently achieves the state-of-the-art retrieval results.

Although the related work achieves breakthrough, there are still two main problems worthy of attention. First, most previous methods including UDCMH preserve the original neighborhood relations from different modalities respectively, while the similarity information from different views

[†] Equal contribution
[*] Corresponding author

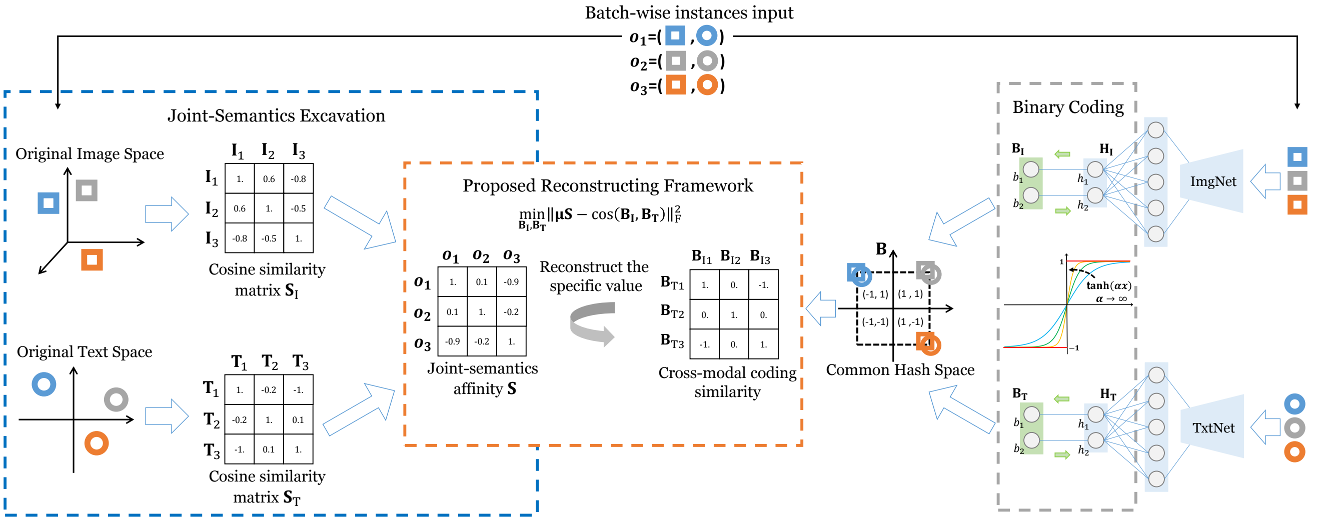

Batch-wise instances input
o_1=(□ , ○)
o_2=(□ , ○)
o_3=(□ , ○)
Joint-Semantics Excavation
Original Image Space
$\mathbf{I}_1$ $\mathbf{I}_2$ $\mathbf{I}_3$
$\mathbf{I}_1$ 1. 0.6 -0.8
$\mathbf{I}_2$ 0.6 1. -0.5
$\mathbf{I}_3$ -0.8 -0.5 1.
Cosine similarity matrix $\mathbf{S}_\mathrm{I}$
Original Text Space
$\mathbf{T}_1$ $\mathbf{T}_2$ $\mathbf{T}_3$
$\mathbf{T}_1$ 1. -0.2 -1.
$\mathbf{T}_2$ -0.2 1. 0.1
$\mathbf{T}_3$ -1. 0.1 1.
Cosine similarity matrix $\mathbf{S}_\mathrm{T}$
Proposed Reconstructing Framework
$\min_{\mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{T}} \|\boldsymbol{\mu}\mathbf{S} - \cos(\mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{T})\|_F^2$
o_1 o_2 o_3
o_1 1. 0.1 -0.9
o_2 0.1 1. -0.2
o_3 -0.9 -0.2 1.
Joint-semantics affinity $\mathbf{S}$
Reconstruct the specific value
$\mathbf{B}_{\mathrm{I}1}$ $\mathbf{B}_{\mathrm{I}2}$ $\mathbf{B}_{\mathrm{I}3}$
$\mathbf{B}_{\mathrm{T}1}$ 1. 0. -1.
$\mathbf{B}_{\mathrm{T}2}$ 0. 1. 0.
$\mathbf{B}_{\mathrm{T}3}$ -1. 0. 1.
Cross-modal coding similarity
$\mathbf{B}$
(-1, 1) (1 , 1)
(-1,-1) (1 ,-1)
Common Hash Space
Binary Coding
$\mathbf{B}_\mathrm{I}$
$\mathbf{H}_\mathrm{I}$
b_1 b_2
h_1 h_2
ImgNet
1
−1
$\tanh(\alpha x)$
$\alpha \to \infty$
$\mathbf{B}_\mathrm{T}$
$\mathbf{H}_\mathrm{T}$
b_1 b_2
h_1 h_2
TxtNet

Figure 1. The pipeline of DJSRH, showing with three multi-modal instances $\boldsymbol{o}_k = \lceil \mathbf{I}_k, \mathbf{T}_k \rfloor$ as input. *Best viewed in color.*

are generally complementary to each other and fusing them in advance can bring more precise neighborhood description. Second, the common Laplacian constraint, whose original form is $\sum_{ij} w_{ij} \|\mathbf{B}_i - \mathbf{B}_j\|^2$ with w_{ij} indicating the original similarity between sample i and j while $\mathbf{B}_i$ and $\mathbf{B}_j$ are the to-be-learned binary codes, preserves the original neighborhood structure in a weighted constraint manner. It merely preserves the similarity order according to the magnitude of w_{ij} which is extremely sensitive to the sample composition in each random sampling training batch. E.g., suppose picking a batch of samples that are dissimilar to each other, and although the similarity weight w among them are all small, the $\mathbf{B}_i$ and $\mathbf{B}_j$ that corresponding to the largest w_{ij} will still be constrained to have too similar or even the same hash codes since their w_{ij} is the relatively largest one in current batch, whereas it is improper from the holistic adjacent structure. Thus, the relevant methods need to first learn the whole training samples' binary codes $\mathbf{B}$ simultaneously, using the complete affinity matrix to ensure the accuracy, and subsequently train the deep network to map the original data to the learned $\mathbf{B}$ in a belated batch-wise manner. Obviously, it not only brings extreme time and space complexities in the first step since employing the complete affinity matrix, but also restrains the interaction between the deep network learning and hash coding part which has been demonstrated to play an essential role in the performance of deep hashing network [17, 19].

In light of these issues, we propose a better unsupervised cross-modal hashing method termed **D**eep **J**oint-**S**emantics **R**econstructing **H**ashing (DJSRH), which has the following main contributions:

- To the best of our knowledge, DJSRH is the first work in deep cross-modal hashing to propose a joint-semantics affinity matrix for the input multi-modal instances, which elaborately integrates the original neighborhood relations from different modalities and accordingly is capable to capture the latent intrinsic semantic affinity among the instances.
- Later DJSRH trains the deep networks to generate binary codes that maximally reconstruct above joint-semantics relations via the proposed reconstructing framework. On the one hand, it adds a linear transformation for the original similarity range to regulate a superior quantization area. On the other hand, it impels to reconstruct the specific similarity value instead of merely preserving the similarity order, which is more competent for the batch-wise training than the Laplacian constraint manner.
- Extensive experiments exhibit the significant improvement by DJSRH in unsupervised cross-modal retrieval and detailed demonstrations for the effectiveness of each component are also provided.

The pipeline of DJSRH is shown in Figure 1, and the rest of this paper is organized as follows. Section 2 briefly introduces the related methods while Section 3 presents our proposed algorithm in detail. Comprehensive experiments on cross-modal retrieval are given in Section 4 and Section 5 makes a summary of this paper.

2. Related Work

In this section, we briefly review some representative unsupervised cross-modal hashing methods which can be roughly categorized into the shallow and the deep schemes according to whether they use the deep networks.

As the earlier shallow approaches, both Cross-View Hashing (CVH) [16] and Inter-Media Hashing (IMH) [25] can be regarded as extending Spectral Hashing [29] to the multi-modal scenario. Collective Matrix Factorization Hashing (CMFH) [6] learns unified hash codes via the collective matrix factorization with a latent factor model for different modalities data. Latent Semantic Sparse Hashing (LSSH) [37] respectively utilizes the sparse coding and the matrix factorization to extract the latent features for images and texts, which later are mapped to a common space and quantized to the unified binary codes.

However, above shallow methods cannot explore the complex nonlinear correlations across different modalities, while the deep schemes [31, 10, 35, 11] have shown their superior ability to bridge the modality gap with the high nonlinearity of deep neural networks. Concretely, [28, 26, 11] employ the autoencoder framework to explore the cross-modal reconstruction, generating the unified latent binary codes for the heterogeneous data. [10, 35] train the networks using the adversarial learning [9, 27], which try to capture the feature distribution of different modalities and narrow their modality gap in a minimax game manner. UDCMH [31] combines the matrix factorization and the Laplacian constraint into the network training, explicitly constraining the hash codes to preserve the neighborhood structure of the original data and consequently achieving the state-of-the-art retrieval results. Although these methods make great progress, there is still much room to improve in this area.

3. Joint-Semantics Reconstructing Hashing

We first introduce some definitions that would be used later. As our method focus on the batch-wise training, the variables will be expressed in the batch manner. Specifically, we use m to denote the batch size and $\boldsymbol{\mathcal{O}} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_m\}$ to represent the m instances in each batch, with each instance is described by a co-occurred image-text pair $\boldsymbol{o}_k = \lceil \mathbf{I}_k, \mathbf{T}_k \rfloor$ in our concerned problem. For each random sampling training batch $\{\boldsymbol{o}_k = \lceil \mathbf{I}_k, \mathbf{T}_k \rfloor\}_{k=1}^m$, we use $\mathbf{F}_\mathrm{I} \in \mathbb{R}^{m \times p_\mathrm{I}}$ and $\mathbf{F}_\mathrm{T} \in \mathbb{R}^{m \times p_\mathrm{T}}$ to denote the original given features from the dataset for $\mathbf{I}_k$ and $\mathbf{T}_k$, while $\mathbf{B}_\mathrm{I} \in \{-1, +1\}^{m \times d}$ and $\mathbf{B}_\mathrm{T} \in \{-1, +1\}^{m \times d}$ are the generated binary codes by our ImgNet for $\mathbf{I}_k$ and TxtNet for $\mathbf{T}_k$. d denotes the encoding length.

Moreover, after normalizing $\mathbf{F}_\mathrm{I}, \mathbf{F}_\mathrm{T}$ to $\hat{\mathbf{F}}_\mathrm{I}, \hat{\mathbf{F}}_\mathrm{T}$ which has unit ℓ_2-norm each row, we can calculate the cosine similarity matrices $\mathbf{S}_\mathrm{I} = \hat{\mathbf{F}}_\mathrm{I}\hat{\mathbf{F}}_\mathrm{I}^\top \in [-1, +1]^{m \times m}$ and $\mathbf{S}_\mathrm{T} = \hat{\mathbf{F}}_\mathrm{T}\hat{\mathbf{F}}_\mathrm{T}^\top \in [-1, +1]^{m \times m}$ to describe the original neighborhood structure for the input images and texts respectively. Besides, as is shown in Figure 1, we can regard the generated binary codes $\mathbf{B}_\mathrm{I}$ and $\mathbf{B}_\mathrm{T}$ as the feature vectors which can only take the vertices of a hypercube. From this perspective, the adjacent vertices correspond to the similar hash codes, that is, the Hamming distance between two binary codes can be indicated by their angular distance. Thus, to describe the neighborhood structure in the Hamming space, we calculate the pairwise cosine similarity matrix $\cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T}) \in [-1, +1]^{m \times m}$ with each element calculating the mutual cosine relation between the hash codes of Image i and Text j: $\cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T})_{ij} = \frac{\mathbf{B}_{\mathrm{I}i}\mathbf{B}_{\mathrm{T}j}^\top}{\|\mathbf{B}_{\mathrm{I}i}\|_2 \|\mathbf{B}_{\mathrm{T}j}\|_2} \in [-1, +1]$. $\mathbf{B}_{\mathrm{I}i}$ indicates the i-th row in $\mathbf{B}_\mathrm{I}$ and $\mathbf{B}_{\mathrm{T}j}$ indicates the j-th row in $\mathbf{B}_\mathrm{T}$. Such a cosine matrix reflects the angular relations among the generated binary codes, which are equivalent to their Hamming distance relations as discussed above.

3.1. Constructing Joint-Semantics Matrix

As demonstrated in [31, 23], learning binary codes that preserve the neighborhood structure of the original data is an effective improvement for the unsupervised training of deep hashing network. To be specific, for cross-modal retrieval task, given the batch-wise input instances $\{\boldsymbol{o}_k = \lceil \mathbf{I}_k, \mathbf{T}_k \rfloor\}_{k=1}^m$ with $\mathbf{F}_\mathrm{I}$ and $\mathbf{F}_\mathrm{T}$, we can calculate the cosine similarity matrices $\mathbf{S}_\mathrm{I} \in [-1, +1]^{m \times m}$ on $\mathbf{F}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T} \in [-1, +1]^{m \times m}$ on $\mathbf{F}_\mathrm{T}$ to describe the original affinity structure in different modalities, and subsequently employ both these two similarity matrices to guide the hash codes learning for $\mathbf{I}_k$ and $\mathbf{T}_k$. Particularly, how to involve $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$ in the training process occupies an important position to the algorithm performance. Most previous shallow or deep cross-modal hashing methods simply preserve these two affinity matrices in a separated manner, which has the following common formulation:

$$\min_{\mathbf{B}} \beta \mathrm{Tr}(\mathbf{B}^\top \mathbf{L}_\mathrm{I} \mathbf{B}) + (1 - \beta)\mathrm{Tr}(\mathbf{B}^\top \mathbf{L}_\mathrm{T} \mathbf{B}), \quad s.t.\ \mathbf{B} \in \{-1, +1\}^{m \times d}, \tag{1}$$

where $\mathbf{L}_\mathrm{I} = \mathrm{diag}(\mathbf{S}_\mathrm{I}\mathbf{1}) - \mathbf{S}_\mathrm{I}$ and $\mathbf{L}_\mathrm{T} = \mathrm{diag}(\mathbf{S}_\mathrm{T}\mathbf{1}) - \mathbf{S}_\mathrm{T}$ are the graph Laplacian matrices. $\beta \in [0, 1]$ is the trade-off parameter that regulates the importance of the neighborhood information from different modalities.

For above Equation (1), on the one hand, the Laplacian constraint manner is unsuitable to the batch-wise network training which we will elaborate in next section. On the other hand, it uses two terms to individually preserve the neighborhood structure $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$ under a co-training pattern, which is suboptimal since the similarity matrices from different views are generally complementary to each other and carefully integrating them in advance can generally obtain a more accurate neighborhood description.

Therefore, we propose a joint-semantics affinity matrix $\mathbf{S} = \mathcal{C}(\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{T}) \in [-1, +1]^{m \times m}$ to integrate the neighborhood information in $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$, with each $\mathbf{S}_{ij} \in [-1, +1]$ indicates the captured latent semantic similarity between the input instances $\boldsymbol{o}_i$ and $\boldsymbol{o}_j$. To introduce the combination function $\mathcal{C}$, we first merge $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$ with a weighted summation manner as follows:

$$\tilde{\mathbf{S}} = \beta \mathbf{S}_{\mathrm{I}} + (1-\beta)\mathbf{S}_{\mathrm{T}}, \quad \beta \in [0,1]. \tag{2}$$

Next, we regard each row in $\tilde{\mathbf{S}} \in [-1,+1]^{m\times m}$ as a new feature for each instance which records the similarity relations between this instance with others, and then calculate the $\tilde{\mathbf{S}}\tilde{\mathbf{S}}^{\top}$ to achieve a high order neighborhood description based on the principle that two semantic relevant instances should share the same similarity relations with other instances. That is, the result of the dot product between their respective row in $\tilde{\mathbf{S}}$ should take a large value. Therefore, we finally employ:

$$\begin{aligned} \mathbf{S} =\ & \mathcal{C}(\mathbf{S}_{\mathrm{I}}, \mathbf{S}_{\mathrm{T}}) \\ =\ & (1-\eta)\ \tilde{\mathbf{S}} + \eta\ \frac{\tilde{\mathbf{S}}\tilde{\mathbf{S}}^{\top}}{m} \\ =\ & (1-\eta)\ [\ \beta\mathbf{S}_{\mathrm{I}} + (1-\beta)\mathbf{S}_{\mathrm{T}}\ ]\ +\ \frac{\eta}{m}\ [\ \beta^2\mathbf{S}_{\mathrm{I}}\mathbf{S}_{\mathrm{I}}^{\top} + \\ & \beta(1-\beta)\mathbf{S}_{\mathrm{I}}\mathbf{S}_{\mathrm{T}}^{\top} + \beta(1-\beta)\mathbf{S}_{\mathrm{T}}\mathbf{S}_{\mathrm{I}}^{\top} + (1-\beta)^2\mathbf{S}_{\mathrm{T}}\mathbf{S}_{\mathrm{T}}^{\top}\ ] \end{aligned} \tag{3}$$

to combine the original neighborhood information $\mathbf{S}_{\mathrm{I}}$ and $\mathbf{S}_{\mathrm{T}}$ from different modalities. Dividing the batch size m is for normalizing $\frac{\tilde{\mathbf{S}}\tilde{\mathbf{S}}^{\top}}{m} \in [-1,+1]^{m\times m}$ and η is the trade-off parameter to adjust the importance of the high order neighborhood description.

Compared with the individual co-training manner (1), Equation (3) combines the affinity information across different modalities in a more explicit and advanced manner. The joint matrix $\mathbf{S} \in [-1,+1]^{m\times m}$ refines the affinity relations from different views ($\mathbf{S}_{\mathrm{I}}$, $\mathbf{S}_{\mathrm{T}}$ and the high order neighborhood description $\tilde{\mathbf{S}}\tilde{\mathbf{S}}^{\top}$) which makes it highly competent to capture the latent intrinsic semantic affinity among the input instances. As a result, we can later learn semantic relevant binary codes for different modalities data with above joint-semantics matrix $\mathbf{S}$ as the self-supervised signal. It greatly helps to learn consistent representations and accordingly improves the retrieval performance.

By the way, it is interesting to notice that our proposed combination function, $\mathbf{S} = (1-\eta)\ \tilde{\mathbf{S}} + \eta\ \frac{\tilde{\mathbf{S}}\tilde{\mathbf{S}}^{\top}}{m}$, conforms to the definition of the diffusion processes [7]. The proposed combination can be seen as taking only one diffusion step for the affinity matrix $\tilde{\mathbf{S}}$ with following update scheme: $\mathbf{W}_{t+1} = \eta\mathbf{W}_t\mathbf{T} + (1-\eta)\mathbf{Y}$, where $\mathbf{W}_0 = \mathbf{Y} = \tilde{\mathbf{S}}$ is the initial affinity matrix, $\mathbf{T} = \frac{\tilde{\mathbf{S}}^{\top}}{m}$ is the transition matrix, and t indicates the t-th step. That is, we merge the original neighborhood matrices $\mathbf{S}_{\mathrm{I}}$ and $\mathbf{S}_{\mathrm{T}}$ to $\tilde{\mathbf{S}}$ and then take a diffusion step on $\tilde{\mathbf{S}}$ with above update scheme to form our eventual $\mathbf{S} = \mathbf{W}_1$. Therefore, [7] provides the other perspective to demonstrate the efficacy of our proposed combination pattern (3). It is desirable to explore the performance when taking more steps or using other diffusion schemes introduced in [7], and we leave them as our future work.

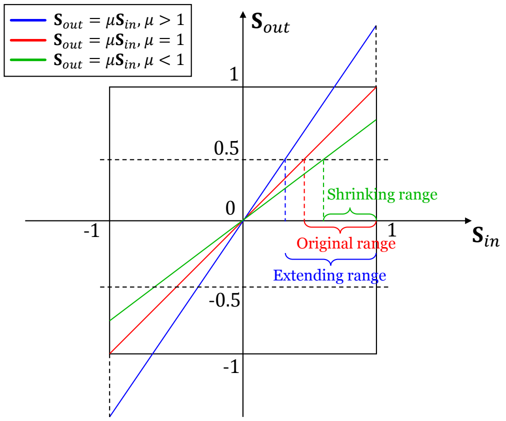


Figure 2. Adding μ to regulate the quantization area for $\mathbf{S}$.

3.2. Reconstructing with Binary Codes

In the last subsection, we have constructed the joint-semantics affinity matrix $\mathbf{S}$ to excavate the latent semantic relations for the batch-wise input instances. Now we can learn semantic relevant binary codes via minimizing the reconstruction error between the desired neighborhood matrix $\mathbf{S}$ and the to-be-learned hash codes structure $\cos(\mathbf{B}_{\mathrm{I}}, \mathbf{B}_{\mathrm{T}})$[1] with the following formulation:

$$\begin{aligned} \min_{\mathbf{B}_{\mathrm{I}}, \mathbf{B}_{\mathrm{T}}}\ & \|\ \mu\mathbf{S} - \cos(\mathbf{B}_{\mathrm{I}}, \mathbf{B}_{\mathrm{T}})\ \|_{\mathrm{F}}^2, \\ s.t.\ & \mathbf{S} = \mathcal{C}(\mathbf{S}_{\mathrm{I}}, \mathbf{S}_{\mathrm{T}}) \in [-1,+1]^{m\times m}. \end{aligned} \tag{4}$$

There are two highlights in the proposed reconstructing framework (4). The first one is adding the hyper-parameter μ which makes our reconstruction more flexible, while the second is reconstructing the specific similarity value which is more compatible with the batch-wise training than the Laplacian constraint pattern.

We analyze the effect of μ at first. Here we take the case of 2-bits hash coding to illustrate the insight. In this 2-bits situation, the hash codes can only take the positions of $(+1,+1), (+1,-1), (-1,+1)$ and $(-1,-1)$ while their mutual cosine similarity can only take '-1', '0' and '$+1$' relations. As we desire to maximally reconstruct the joint-semantics structure $\mathbf{S} \in [-1,+1]^{m\times m}$ with these 2-bits hash codes, the original similarity range $[0.5, 1]$ in $\mathbf{S}$ will be assigned to '$+1$' relation in the Hamming space, i.e., the corresponding image-text pairs will be impelled to take the same binary codes. Similarly, $(-0.5, 0.5)$ will be assigned to '0' and $[-1,-0.5]$ to '-1'. However, above quantization process is too stiff to learn reasonable hash codes. E.g., a semantic relevant image-text pair is totally possible to take 0.4 similarity value in the captured $\mathbf{S}$ whereas it will

[1] $\cos(\mathbf{B}_{\mathrm{I}}, \mathbf{B}_{\mathrm{T}}) \in [-1,+1]^{m\times m}$ reflects the current hash codes neighborhood structure in the Hamming space as discussed in the beginning of Section 3.

be toughly quantized to the closest ‘ 0 ’ instead of the better ‘$+1$’ to share the same binary codes. Regarding this deficiency, as is shown in Figure 2, we add a hyper-parameter μ to realize a linear transformation for the original similarity matrix $\mathbf{S}$ which adjusts the corresponding similarity range for the limited relations in the Hamming space. Taking relation ‘$+1$’ as example, $\mu > 1$ means extending the original range $[0.5, 1]$, enabling more image-text pairs to be quantized with ‘$+1$’ relation and accordingly in possession of the same hash codes, while $\mu < 1$ means shrinking the ‘$+1$’ range inversely. Thus, the parameter μ in the proposed framework (4) helps to regulate a superior quantization area for $\mathbf{S}$ which highly improves the flexibility of our reconstruction.

Next, we analyze the superiority of framework (4) in the batch-wise training. Based on the discussion in Introduction, the widely used Laplacian constraint scheme $\mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) = \sum_{i,j}\mathbf{S}_{ij}\|\mathbf{B}_i - \mathbf{B}_j\|^2$ merely constrains the binary codes to preserve the original similarity order in a weighted constraint manner, i.e., if $\mathbf{S}_{12} > \mathbf{S}_{13}$ then $\mathbf{B}_1$ should be more similar to $\mathbf{B}_2$ than $\mathbf{B}_3$, while such relative order will be extremely sensitive to the sample composition in each random sampling training batch. For example, suppose $\mathbf{S}_{12} = 0.2, \mathbf{S}_{13} = 0.1, \mathbf{S}_{23} = 0.1$ in current batch with three samples, then $\mathbf{B}_1$ should be more similar to $\mathbf{B}_2$ than $\mathbf{B}_3$ whereas the specific similarity degree is not defined and it is probable that $\mathbf{B}_1$ and $\mathbf{B}_2$ will be constrained to take excessively similar or even the same hash codes since $\mathbf{S}_{12}$ is the relatively largest one in current batch, which is obviously improper from the holistic adjacent structure as $\mathbf{S}_{12} \neq 1$. Thereby, the relevant methods have to learn binary codes for the whole training samples simultaneously, using the complete $n \times n$ affinity matrix to ensure the algorithm precision which inevitably brings the high time and space complexities in the training stage. In contrast, the proposed reconstructing framework (4) maximally reconstructs the specific similarity value in $\mathbf{S}$ instead of their similarity order. It is insensitive to the composition of each random sampling training batch and accordingly be more competent for the batch-wise input manner, qualifying our coding networks for the desired end-to-end batch-wise training. Compared with the previous Laplacian methods, framework (4) not only dramatically reduces the algorithm complexity but also help to achieve better coding performance due to the increasing interaction between the deep network learning and hash coding part in each batch.

Therefore, we employ framework (4) as the basic pattern to compose our overall training objective. To be concrete, besides the component in framework (4) acting as the inter-modal reconstruction with $\mathbf{B}_\mathrm{I}$ and $\mathbf{B}_\mathrm{T}$, we also supplement the intra-modal reconstruction since considering both intra- and inter-view in the cross-modal network training has been demonstrated for effectively improving the retrieval performance [21, 25, 32]. Thus the final training objective of the proposed DJSRH is:

$$\begin{aligned}\min_{\mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{T}} \|\mu\mathbf{S} - \cos(\mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{T})\|_\mathrm{F}^2 &+ \lambda_1\|\mu\mathbf{S} - \cos(\mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{I})\|_\mathrm{F}^2 \\ &+\lambda_2\|\mu\mathbf{S} - \cos(\mathbf{B}_\mathrm{T},\mathbf{B}_\mathrm{T})\|_\mathrm{F}^2, \\ s.t.\ \mathbf{S} = \mathcal{C}(\mathbf{S}_\mathrm{I},&\mathbf{S}_\mathrm{T}) \in [-1,+1]^{m\times m}, \\ \mathbf{B}_\mathrm{I},\mathbf{B}_\mathrm{T} &\in \{-1,+1\}^{m\times d},\end{aligned} \tag{5}$$

where λ_1 and λ_2 are the trade-off parameters to balance the inter-modal and intra-modal reconstruction. $\mathcal{C}$ is the proposed combination function (3) to integrate $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$.

Algorithm 1 Deep Joint-Semantics Reconstructing Hashing

```
Input:
    Training set $\{\boldsymbol{o}_k = \lceil\mathbf{I}_k, \mathbf{T}_k\rfloor\}_{k=1}^{n}$ and their corresponding original features $\mathbf{F}_\mathrm{I}$ and $\mathbf{F}_\mathrm{T}$; ImgNet $\mathcal{G}_{\theta_\mathrm{I}}$ and TxtNet $\mathcal{G}_{\theta_\mathrm{T}}$ with $\theta_\mathrm{I}$ and $\theta_\mathrm{T}$ denoting the deep network parameters; batch size $m$;
Output:
    Hashing coding function $\varphi_\mathrm{I}(x) = \mathrm{sgn}(\mathcal{G}_{\theta_\mathrm{I}}(x))$ for image input and $\varphi_\mathrm{T}(x) = \mathrm{sgn}(\mathcal{G}_{\theta_\mathrm{T}}(x))$ for text input;
Initialize epoch $t = 0$;
repeat
    $t = t + 1;\ \alpha = \sqrt{t}$;
    for $\lfloor\frac{n}{m}\rfloor$ iterations do
        Randomly sample a batch of instances from training set $\{\boldsymbol{o}_k = \lceil\mathbf{I}_k, \mathbf{T}_k\rfloor\}_{k=1}^{m}$;
        Calculate the normalized $\hat{\mathbf{F}}_\mathrm{I}, \hat{\mathbf{F}}_\mathrm{T}$ and integrate the cosine matrices $\mathbf{S}_\mathrm{I} = \hat{\mathbf{F}}_\mathrm{I}\hat{\mathbf{F}}_\mathrm{I}^\top$, $\mathbf{S}_\mathrm{T} = \hat{\mathbf{F}}_\mathrm{T}\hat{\mathbf{F}}_\mathrm{T}^\top$ to the joint-semantics affinity $\mathbf{S}$ with Equation (3);
        Forward propagate $\mathbf{H}_\mathrm{I} = \mathcal{G}_{\theta_\mathrm{I}}(\mathbf{I}), \mathbf{H}_\mathrm{T} = \mathcal{G}_{\theta_\mathrm{T}}(\mathbf{T})$;
        Hash coding with activation function (7) $\mathbf{B}_\mathrm{I} = \tanh(\alpha\mathbf{H}_\mathrm{I}), \mathbf{B}_\mathrm{T} = \tanh(\alpha\mathbf{H}_\mathrm{T})$;
        Calculate the objective function (5), back propagate the gradients with the chain rule and update the whole parameters;
    end for
until convergence
```

3.3. Optimization

The major difficulty to optimize the objective function (5) lies in the discrete constraint imposed on the binary code $\mathbf{B}_\mathrm{I}$ and $\mathbf{B}_\mathrm{T}$. For deep hashing network, if we denote the output of the last hidden layer (without activation function) as $\mathbf{H} \in \mathbb{R}^{m\times d}$ (represents both $\mathbf{H}_\mathrm{I}$ in ImgNet and $\mathbf{H}_\mathrm{T}$ in TxtNet), we can generate the strict binary hash codes by:

$$\mathbf{B} = \mathrm{sgn}(\mathbf{H}) \in \{-1,+1\}^{m\times d}, \tag{6}$$

where $\mathrm{sgn}(\cdot)$ is the sign function that outputs +1 for positive input and -1 otherwise on each element.

However in the backward propagation, the gradient of the sign function is zero for all nonzero input, which will ruinously block the gradients back to the front layers. To handle this vanishing gradients problem, we follow [4, 11] to adopt a scaled tanh function:

$$\mathbf{B} = \tanh(\alpha\mathbf{H}) \in [-1, +1]^{m\times d},\ \alpha \in \mathbb{R}^{+} \qquad (7)$$

with increasing α during the training stage, to replace the encoding function (6). It is motivated by a key observation that $\lim_{\alpha\to\infty}\tanh(\alpha x) = \mathrm{sgn}(x)$, as is shown in the right part of Figure 1. Accordingly, the tightening tanh function generates a sequence of smoothed optimization problems which with increasing α can converge to the original intractable binary coding problem (5).

The overall procedure of our proposed DJSRH is summarized in Algorithm 1.

4. Experiments

The source codes are available at: `https://github.com/zzs1994/DJSRH`.

4.1. Datasets

Wiki [22] consists of 2,866 multimedia documents in 10 categories from Wikipedia. Each document serves as an instance containing one image and text with at least 70 words. A hand-crafted 128-dimensional SIFT feature vector is also provided for each image, while each text is accompanied with a 10-dimensional topic vector generated by the Latent Dirichlet Allocation (LDA) model.

NUS-WIDE [5] contains 269,648 multi-modal instances, each of which consists of an image and the associated textual tags. Following previous methods, the top 10 most frequent labels from the original 81 classes are selected and the corresponding 186,577 annotated instances are preserved. A 500-dimensional BoW SIFT feature is provided for each image while an index vector of the most frequent 1,000 textual tags (a.k.a the tag occurrence feature) is sorted out for each text.

MIRFlickr [13] composes of 25,000 instances annotated with 24 provided labels, with each instance containing an image and the associated textual tags. SIFT descriptor is provided for each image and tag occurrence feature is sorted out for each text.

4.2. Evaluation Criterion

Wiki is officially split into the database and the query set with 2,173 and 693 instances respectively. As for MIRFlickr and NUS-WIDE, following [31, 6] 2,000 instances are randomly picked out as the query while the rest as the database. Besides, the whole database of Wiki will serve as its training set due to its small size, while for the larger MIRFlickr and NUS-WIDE 5,000 instances are randomly sampled from the database for training. Later in the evaluation step, the trained hash coding function will be applied on each instance in the database and the query set to obtain their final binary representations.

We adopt the two common retrieval metrics: mean Average Precision (mAP) and precision@top-R curve to evaluate the performance of the proposed DJSRH and baselines. Any two data points are considered to be the ground-truth neighbors if they share at least one common label.

4.3. Implementation Details

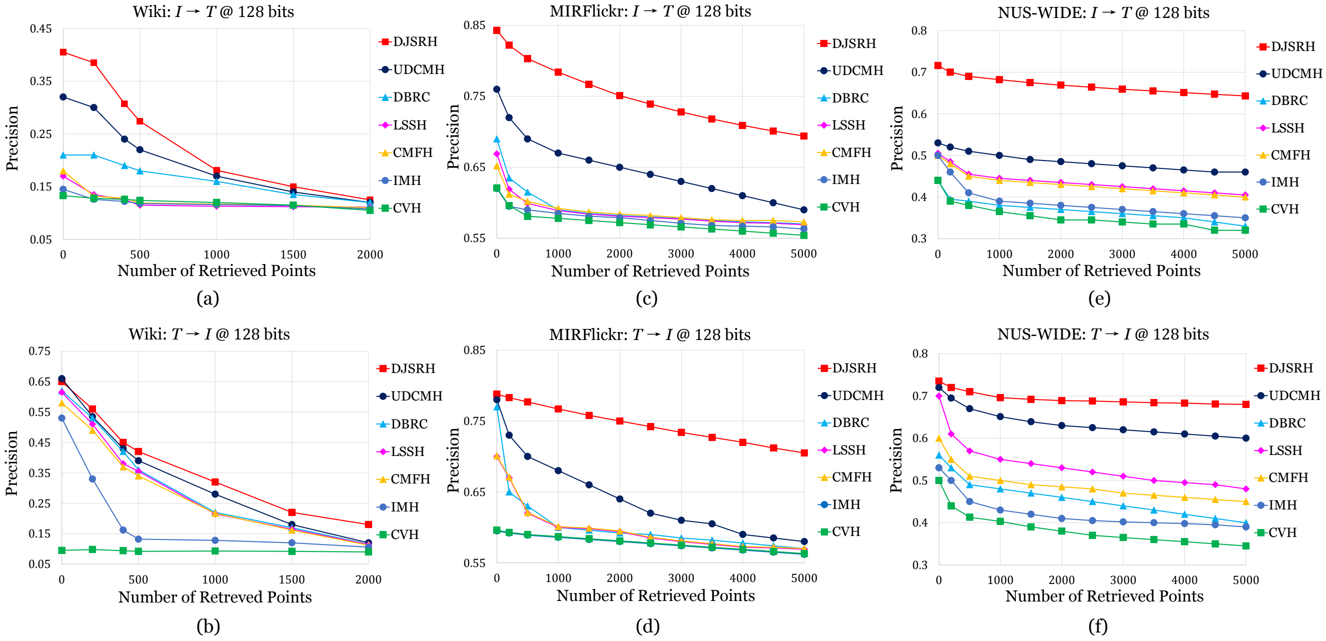
As the hand-crafted SIFT features are insufficient to capture the abstract semantic relations of images, we follow the previous work to extract the deep features from CNN (pretrained on ImageNet) to replace the SIFT descriptors. Specifically, we extract the 4,096-dimensional features from the fc7 layer (after ReLU) of AlexNet [15] as the original image features $\mathbf{F}_{\mathrm{I}} \in \mathbb{R}^{m\times 4096}$ for the batch-input images $\{\mathbf{I}_k\}_{k=1}^{m}$, while for texts $\{\mathbf{T}_k\}_{k=1}^{m}$ we just adopt the original LDA topic vectors or the tag occurrence features as their $\mathbf{F}_{\mathrm{T}}$. Notably we need to preprocess the $\mathbf{S}_{\mathrm{I}}$ with $\mathbf{S}_{\mathrm{I}} \leftarrow 2\,\mathbf{S}_{\mathrm{I}} - 1$ and the same for the $\mathbf{S}_{\mathrm{T}}$ since that the current $\mathbf{F}_{\mathrm{I}}, \mathbf{F}_{\mathrm{T}}$ are all taking nonnegative numbers and their generated $\mathbf{S}_{\mathrm{I}}, \mathbf{S}_{\mathrm{T}} \in [0, 1]^{m\times m}$ will inevitably lead to some unchangeable hash bits as the smallest similarity is 0 right now. Transforming them back to $[-1, 1]^{m\times m}$ beforehand can prevent this issue.

For fair comparisons, we follow [31] to adopt AlexNet and Multilayer Perceptron (MLP) as the backbone of our ImgNet and TxtNet respectively. We replace the classifier layer fc8 of AlexNet with a new fc of d hidden units to generate the continuous $\mathbf{H}_{\mathrm{I}} \in \mathbb{R}^{m\times d}$, and then obtain $\mathbf{B}_{\mathrm{I}}$ through the coding formula (7) for training and formula (6) for test. For text modality, as the original text descriptions are diverse and difficult to handle, we follow the previous schemes to directly adopt the topic vectors or the tag occurrence features as the input to MLP, i.e., $\mathbf{F}_{\mathrm{T}}$ serves as $\mathbf{T}$ to feed in the TxtNet. The first fc layer in our MLP has 4096 units with the ReLU as their activation function. The second fc has d units to produce $\mathbf{H}_{\mathrm{T}} \in \mathbb{R}^{m\times d}$ which subsequently generates the $\mathbf{B}_{\mathrm{T}}$ through formula (7) for training and formula (6) for test.

Additionally, we fix the batch size as 32 and employ the SGD optimizer with 0.9 momentum and 0.0005 weight decay. We cross-validate the hyper-parameters and finally take $\eta = 0.4$, $\mu = 1.5$ for all three datasets, $\beta = 0.6$, $\lambda_1 = \lambda_2 = 0.1$ for NUS-WIDE, $\beta = 0.9$, $\lambda_1 = \lambda_2 = 0.1$ for MIRFlickr and $\beta = 0.3$, $\lambda_1 = \lambda_2 = 0.3$ for Wiki. Moreover, the learning rates are set to 0.001 for the ImgNet and 0.01 for the TxtNet when running on NUS-WIDE and MIRFlickr. As for Wiki which contains much fewer instances, we fix the convolutional layers of the ImgNet with the pretrained parameters and only update the fully connected layers, setting

Task	Method	Wiki				MIRFlickr				NUS-WIDE			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
$I \rightarrow T$	CVH	0.179	0.162	0.153	0.149	0.606	0.599	0.596	0.598	0.372	0.362	0.406	0.390
	IMH	0.201	0.203	0.204	0.195	0.612	0.601	0.592	0.579	0.470	0.473	0.476	0.459
	CMFH	0.251	0.253	0.259	0.263	0.621	0.624	0.625	0.627	0.455	0.459	0.465	0.467
	LSSH	0.197	0.208	0.199	0.195	0.584	0.599	0.602	0.614	0.481	0.489	0.507	0.507
	DBRC	0.253	0.265	0.269	0.288	0.617	0.619	0.620	0.621	0.424	0.459	0.447	0.447
	UDCMH	0.309	0.318	0.329	0.346	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558
	DJSRH	**0.388**	**0.403**	**0.412**	**0.421**	**0.810**	**0.843**	**0.862**	**0.876**	**0.724**	**0.773**	**0.798**	**0.817**
$T \rightarrow I$	CVH	0.252	0.235	0.171	0.154	0.591	0.583	0.576	0.576	0.401	0.384	0.442	0.432
	IMH	0.467	0.478	0.453	0.456	0.603	0.595	0.589	0.580	0.478	0.483	0.472	0.462
	CMFH	0.595	0.601	0.616	0.622	0.642	0.662	0.676	0.685	0.529	0.577	0.614	0.645
	LSSH	0.569	0.593	0.593	0.595	0.637	0.659	0.659	0.672	0.577	0.617	0.642	0.663
	DBRC	0.574	0.588	0.598	0.599	0.618	0.626	0.626	0.628	0.455	0.459	0.468	0.473
	UDCMH	**0.622**	0.633	0.645	**0.658**	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716
	DJSRH	0.611	**0.635**	**0.646**	**0.658**	**0.786**	**0.822**	**0.835**	**0.847**	**0.712**	**0.744**	**0.771**	**0.789**

Table 1. The mAP@50 results on image query text ($I \rightarrow T$) and text query image ($T \rightarrow I$) retrieval tasks at various encoding lengths and datasets. The best performances are shown as **Red** while the suboptimal as Blue.



(a) (c) (e)

(b) (d) (f)

Figure 3. The precision@top-R curves on different datasets at 128 encoding length.

0.01 learning rate for both the ImgNet and TxtNet.

4.4. Retrieval Performance

We compare our proposed DJSRH with several representative baselines including CVH [16], IMH [25], CMFH [6], LSSH [37], DBRC [11] and UDCMH [31], in which the former four ones are shallow methods while DBRC, UDCMH and ours are deep schemes.

We first compare the mAP results with the baselines and we follow [31, 6] to set the retrieved points as 50 (i.e., mAP@50). The results are shown in Table 1. As can be seen, the proposed DJSRH significantly outperforms the state-of-the-art unsupervised cross-modal hashing methods at various encoding lengths and datasets. To be specific, compared to the shallow methods which have also used the deep features as their image modality representations, the deep network baselines achieve better results due to they can back propagate the gradients to the front network to learn more complex and competent hash coding function. DJSRH further improves the performance of the deep schemes with the help of the proposed joint-semantics affinity matrix $\mathbf{S}$ and the hash coding framework (4). For a quantitative comparison, we achieve about 15% improvements for $I \rightarrow T$ retrieval (image query text) and 10% improvements for $T \rightarrow I$ (text query image) on both MIRFlickr and NUS-WIDE, while the improvements on Wiki are relative lower (about 8% increase for $I \rightarrow T$ but nearly standing still for $T \rightarrow I$). The main reason is that Wiki contains much fewer instances than other two datasets, which dramatically limits the learning capability of the deep neural networks as is known to all.

Moreover, Figure 3 shows the precision@top-R curves among the compared methods, in which DJSRH still significantly outperforms the state-of-the-art baselines on various datasets which confirms the superiority of our proposed scheme in unsupervised cross-modal retrieval.

4.5. Ablation Study

To further demonstrate the effectiveness of each part in DJSRH, we design several variants to evaluate the performance when adding the proposed components one by one. Following the introduction order in Section 3, **DJSRH-1** and **DJSRH-2** are the basic variants which respectively employ $\|\mathbf{S}_\mathrm{I} - \cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T})\|_\mathrm{F}^2$ and $\|\mathbf{S}_\mathrm{T} - \cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T})\|_\mathrm{F}^2$ as their training objective. **DJSRH-3** is the variant that simply merges the affinity matrices from different modalities with the weighted summation, $\mathbf{S} = \beta\mathbf{S}_\mathrm{I} + (1-\beta)\mathbf{S}_\mathrm{T}$, and then utilizes $\|\mathbf{S} - \cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T})\|_\mathrm{F}^2$ as its training objective. **DJSRH-4** is the variant based on DJSRH-3 which further supplements the high order neighborhood information to improve the joint affinity matrix, i.e., employing Equation (3) to generate $\mathbf{S}$. To go a step further, **DJSRH-5** is the variant adding the regulation parameter μ, namely $\|\mu\mathbf{S} - \cos(\mathbf{B}_\mathrm{I}, \mathbf{B}_\mathrm{T})\|_\mathrm{F}^2$. Then, adding the intra-modal reconstruction terms ($\lambda_1 = \lambda_2 = 0.1$) to DJSRH-5 finally composes our proposed **DJSRH**.

We also set a variant **DJSRH-6** in the end which employs the constant tanh function (i.e., $\alpha = 1$) as the last coding function for the ImgNet and TxtNet, replacing the tightening tanh (7) adopted by DJSRH. The mAP@50 results of all variants are shown in Table 2.

From the table we can observe that each of our proposed components plays a certain role for our final results. Specifically, compared with the performance of the variants DJSRH-1 and DJSRH-2, the incremental precision of DJSRH-3 and DJSRH-4 demonstrate the effectiveness of the proposed combination function (3). Both the modalities mergence (DJSRH-3) and the high order neighborhood information (DJSRH-4) help to refine the original similarities $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{T}$. They can better capture the latent semantic relations, impelling to learn more consistent hash codes and accordingly achieving higher retrieval results. Then, the variant DJSRH-5 and DJSRH exhibit the effect of the regulation parameter μ and the intra-modal reconstruction. DJSRH outperforms the variant DJSRH-6 demonstrating that the tightening tanh can effectively reduce the quantization error caused by the constant tanh as discussed in [4, 11].

Model	Configuration	64bits		128bits	
		$I \rightarrow T$	$T \rightarrow I$	$I \rightarrow T$	$T \rightarrow I$
DJSRH-1	$\mathbf{S} = \mathbf{S}_\mathrm{I}$	0.717	0.712	0.741	0.735
DJSRH-2	$\mathbf{S} = \mathbf{S}_\mathrm{T}$	0.702	0.606	0.734	0.581
DJSRH-3	$\beta\mathbf{S}_\mathrm{I} + (1-\beta)\mathbf{S}_\mathrm{T}$	0.724	0.720	0.747	0.738
DJSRH-4	$+(\eta = 0.4)$	0.790	0.745	0.803	0.757
DJSRH-5	$+(\mu = 1.5)$	0.793	0.747	0.812	0.768
DJSRH	$+(\lambda_1 = \lambda_2 = 0.1)$	**0.798**	**0.771**	**0.817**	**0.789**
DJSRH-6	$-(\alpha = 1)$	0.786	0.770	0.811	0.782

Table 2. The mAP@50 results on NUS-WIDE to evaluate the effectiveness of each component in DJSRH.

Last but not least, we would like to highlight that the variants DJSRH-1,2,3 have surpassed UDCMH (the state-of-the-art previous method in Table 1) which exactly attributes to the superiority of our hash coding framework (4). It facilitates our deep hashing networks for the end-to-end batch-wise training which largely increases the interaction between the deep network learning and hash coding part than the previous Laplacian constraint pattern.

5. Conclusion

In this paper, we propose **D**eep **J**oint-**S**emantics **R**econstructing **H**ashing (DJSRH) for large-scale unsupervised cross-modal retrieval. DJSRH first explicitly integrates the original neighborhood information from different modalities into a joint-semantics affinity matrix, to excavate the latent intrinsic semantic relations among the input instances. Then it learns binary codes to maximally reconstruct above joint-semantics structure via the proposed reconstructing framework, which on the one hand adds a linear transformation for the original similarity range to regulate a better quantization area, making our reconstruction more flexible. On the other hand, it reconstructs the specific similarity value enabling DJSRH to be more competent for the end-to-end batch-wise training than the common Laplacian constraint. Extensive experiments demonstrate the superiority of our proposed method and the effectiveness of each component is also carefully studied.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002400, the National Natural Science Foundation of China under Grant 61671027 and the National Key Basic Research Program of China under Grant 2015CB352303.

References

[1] 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.

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

[3] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *SIGKDD*, pages 1445–1454, 2016.

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. HashNet: Deep learning to hash by continuation. In *ICCV*, pages 5608–5617, 2017.

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *I-CIVR*, pages 48–56, 2009.

[6] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2075–2082, 2014.

[7] Michael Donoser and Horst Bischof. Diffusion processes for retrieval revisited. In *CVPR*, pages 1320–1327, 2013.

[8] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Cross-modal deep variational hashing. In *ICCV*, pages 4077–4085, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing X-u, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

[10] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. Unsupervised cross-modal retrieval through adversarial learning. In *ICME*, pages 1153–1158, 2017.

[11] Di Hu, Feiping Nie, and Xuelong Li. Deep binary reconstruction for cross-modal hashing. *IEEE TMM*, 21(4):973–985, 2018.

[12] Yao Hu, Zhongming Jin, Hongyi Ren, Deng Cai, and Xiaofei He. Iterative multi-view hashing for cross media indexing. In *ACM MM*, pages 527–536, 2014.

[13] Mark J Huiskes and Michael S Lew. The MIR Flickr retrieval evaluation. In *ICMIR*, pages 39–43, 2008.

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

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

[16] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, 2011.

[17] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*, pages 3270–3278, 2015.

[18] 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.

[19] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In *IJCAI*, pages 1711–1717, 2016.

[20] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, pages 3864–3872, 2015.

[21] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *IEEE TPAMI*, 36(4):824–830, 2014.

[22] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE TPAMI*, 36(3):521–535, 2014.

[23] Fumin Shen, Yan Xu, Li Liu, Yang Yang, Zi Huang, and Heng Tao Shen. Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE TPAMI*, 40(12):3034–3044, 2018.

[24] Yuming Shen, Li Liu, Ling Shao, and Jingkuan Song. Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval. In *ICCV*, pages 4097–4106, 2017.

[25] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD*, pages 785–796, 2013.

[26] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *IJCAI*, 2015.

[27] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui X-u, Benyou Wang, Peng Zhang, and Dell Zhang. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. In *SIGIR*, pages 515–524, 2017.

[28] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. *VLDB*, 7(8):649–660, 2014.

[29] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *NeurIPS*, pages 1753–1760, 2009.

[30] Botong Wu, Qiang Yang, Wei-Shi Zheng, Yizhou Wang, and Jingdong Wang. Quantized correlation hashing for fast cross-modal search. In *IJCAI*, 2015.

[31] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In *IJCAI*, pages 2854–2860, 2018.

[32] 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.

[33] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

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

[35] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *AAAI*, 2018.

[36] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *ECCV*, pages 591–606, 2018.

[37] Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *SIGIR*, pages 415–424, 2014.