Task-adaptive Asymmetric Deep Cross-modal Hashing

Tong Wang[a], Lei Zhu[a], Zhiyong Cheng[b], Jingjing Li[c], Huaxiang Zhang[a]

[a] *School of Information Science and Engineering, Shandong Normal University*
[b] *Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences)*
[c] *School of Computer Science and Engineering, University of Electronic Science and Technology of China*

Abstract

Supervised cross-modal hashing aims to embed the semantic correlations of heterogeneous modality data into the binary hash codes with discriminative semantic labels. Because of its advantages on retrieval and storage efficiency, it is widely used for solving efficient cross-modal retrieval. However, existing researches equally handle the different tasks of cross-modal retrieval, and simply learn the same couple of hash functions in a symmetric way for them. Under such circumstance, the uniqueness of different cross-modal retrieval tasks are ignored and sub-optimal performance may be brought. Motivated by this, we present a Task-adaptive Asymmetric Deep Cross-modal Hashing (TA-ADCMH) method in this paper. It can learn task-adaptive hash functions for two sub-retrieval tasks via simultaneous modality representation and asymmetric hash learning. Unlike previous cross-modal hashing approaches, our learning framework jointly optimizes semantic preserving that transforms deep features of multimedia data into binary hash codes, and the semantic regression which directly regresses query modality representation to explicit label. With our model, the binary codes can effectively preserve semantic correlations across different modalities, meanwhile, adaptively capture the query semantics. The superiority of TA-ADCMH is proved on two standard datasets from many aspects.

Keywords: Cross-modal Similarity Retrieval, Task-adaptive, Asymmetric Deep Hashing Learning

Email address: leizhu0608@gmail.com (Lei Zhu)

1. Introduction

Cross-modal retrieval [1, 2, 3, 4, 5, 6, 7] takes a certain kind of modality data as query objects to retrieve the relevant data in other modalities. Meanwhile, a large amounts of heterogeneous multi-modal data are explosively generated in various social networks. To tackle the retrieval efficiency problem, cross-modal hashing [8, 9, 10, 11, 12, 13, 14, 15, 16, 17] is proposed to project the high-dimensional multi-modal data into the low-dimensional binary hash codes, which are forced to express consistent semantics with the original data. For the high retrieval and storage efficiency, it has aroused considerable attention to solve large-scale cross-modal search.

With the trend, the hashing methods in the field of cross-modal search have become a research hotspot in plenty of literatures. There are two main categories of these methods: unsupervised [18, 19, 20, 8, 21, 22] and supervised [23, 24, 15, 11, 25, 13] cross-modal hashing. Unsupervised cross-modal hashing methods learn the low-dimensional embedding of original data without any semantic labels. The generated hash codes are learned to keep the semantic correlation of heterogeneous multi-modal data. Contrastively, supervised cross-modal hashing methods exhibit a strong discrimination capability in the learning process of hash codes with the supervision of explicit semantic labels.

The shallow cross-modal hashing method has always been the core and main force of cross-modal retrieval and achieved promising results. With the problem studied deeply, the biggest defect of these methods is that the hash functions depend on linear or simple nonlinear projections. This may limit the discriminative capability of modality feature representation, and result in the low retrieval accuracy of the learned binary codes. Recently, deep cross-modal hashing [26, 27, 28, 29] is proposed to simultaneously perform deep representation and hash code learning. They replace the linear mapping with multi-layer nonlinear mapping and thus capture the intrinsic semantic correlations of cross-modal instances more effectively. It has been proved that cross-modal hashing methods based on deep models have better performance than the shallow hash models which directly adopt hand-crafted features.

Although great success has been achieved by existing methods, they equally handle the cross-modal retrieval tasks (e.g. image retrieves text and text retrieves image), and simply learn the same couple of hash functions for them. Under such circumstance, the characteristics of different cross-modal retrieval tasks are ignored and sub-optimal performance may be caused accordingly. To tackle the limitation, this paper proposes a Task-adaptive Asymmetric Deep Cross-modal Hashing (TA-DCMH) method to learn task-specific hash functions for each cross-modal sub-retrieval tasks. The major contributions and innovations are stated as follows:

- We propose a new supervised asymmetric hash learning framework based on deep neural networks for large-scale cross-modal search. Two couples of deep hash functions can be learned for different cross-modal retrieval tasks, by performing simultaneous deep feature representation and asymmetric hash learning. For all we know, no similar work has been proposed yet.

- In asymmetric hash learning part, we jointly optimize the semantic preserving of original data from multiple modalities, and the representation capability enhancement of query modality. With such design, the learned hash codes can establish a semantic connection across different modalities, as well as capture the query semantics of the specific cross-modal retrieval task.

- An iterative optimization algorithm is proposed to enable the discreteness of hash codes and alleviate the errors of binary quantization. The results of experiment demonstrate that this algorithm is superior on two datasets widely tested in cross-modal retrieval.

2. Literature review of cross-modal hashing

2.1. Unsupervised Cross-modal Hashing

Unsupervised cross-modal hashing transforms the modality features into the shared hash codes by preserving the original similarities. Representative works include Cross-view Hashing (CVH) [18], Inter-media Hashing (IMH) [19], Linear Cross-modal Hashing (LCMH) [20], Collective Matrix Factorization Hashing (CMFH) [8], Latent Semantic Sparse Hashing (LSSH) [30], Robust and Flexible Discrete Hashing (RFDH)

[31], Cross-modal Discrete Hashing (CMDH) [21] and Collective Reconstructive Embeddings (CRE) [22]. CVH is a typical graph-based hashing method extended from the standard spectral hashing [32]. It minimizes the weighted Hamming distances to transform the original multi-view data into the binary codes. IMH maps heterogeneous multimedia data into hash codes by constructing graphs. It learns the hash functions by linear regression for new instances. Its joint learning scheme can effectively preserve the inter- and intra- modality consistency. LCMH first leverages k-means clustering to represent each training data as k-dimensional vector, and then maps the vector into the to-be-learnt binary codes. CMFH utilizes collective matrix factorization model to transform multimedia data into low dimensional space, then approximates it with hash codes. It also fuses the multi-view information to enhance the search accuracy. LSSH follows similar idea of CMFH. It attempts to learn the latent factor matrix for image structures by sparse coding and text concepts by matrix decomposing. Compared with CMFH, it can better capture high-level semantic correlation for similarity search across different modalities. RFDH first learns the unified hash codes for each training data by employing discrete collaborative matrix factorization. Then, it jointly adopts l2,1-norm and adaptively weight of each modality to enhance the robustness and flexibility of hash codes. CMDH presents a discrete optimization strategy to learn the unified binary codes for multiple modalities. This strategy projects the heterogeneous data into a low-dimensional latent semantic space by using matrix factorization. The latent features are quantified as the hash codes by projection matrix. CRE is proposed to learn unified binary codes and binary mappings for different modalities by collective reconstructive embedding. It simultaneously bridges the semantic gap between heterogeneous data.

2.2. Supervised Cross-modal Hashing

Supervised cross-modal hashing generates the hash codes under the guidance of semantic information. Typical methods include Semantic Correlation Maximization (SCM) [23], Semantics-Preserving Hashing (SePH) [24], Supervised Matrix Factorization Hashing (SMFH) [15], Semantic Topic Multimodal hashing (STMH) [11], Discrete Latent Factor Model based Cross-Modal Hashing (DLFH) [33], Discrete Cross-modal Hashing (DCH) [25] and Label Consistent Matrix Factorization Hashing (LCMFH)

[13]. SCM aims at preserving maximum semantic information into hash codes by avoiding computing pair-wise semantic matrix explicitly. It improves both the retrieval speed and space utilization. SePH first employs probability distribution to preserve supervision information of multi-modal data, and then the hash codes can be obtained by solving the problem of Kullback-Leibler divergence. SMFH is developed based on the collective matrix decomposing. It jointly employs graph Laplacian and semantic label to learn binary codes for multi-modal data. STMH employs semantic modeling to detect different semantic themes for texts and images respectively, and then maps the captured semantic representations into a low-dimensional latent space to obtain hash codes. DLFH proposes an efficient hash learning algorithm based on the discrete latent factor model to directly learn binary hash codes for cross-modal retrieval. DCH is an extended application of Supervised Discrete Hashing (SDH) [34] in multi-modal retrieval. It learns a set of modality-dependence hash projections as well as discriminative binary codes to keep the classification consistent with the label for multi-modal data. LCMFH leverages the auxiliary matrix to project the original multi-modal data to the low-dimensional representation of latent space, and quantizes it with semantic label to the hash codes.

All the above hashing methods are shallow modeling, which imposes linear or non-linear transformations to construct the hash functions. Thus, these methods cannot effectively explore the semantic correlations of heterogeneous multi-modal data.

2.3. Deep Cross-modal Hashing

They basically seek a common binary semantic space via multi-layer nonlinear projection from multiple heterogeneous modalities. State-of-the-art deep cross-modal hashing methods include Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) [26], Deep Binary Reconstruction for Cross-modal Hashing (DBRC) [27], Deep Cross-Modal Hashing (DCMH) [28], Discrete Deep Cross-Modal Hashing (DD-CMH) [29] and Self-supervised adversarial hashing (SSAH) [35]. UGACH is proposed to promote the learning of hash functions by the confrontation between generative model and discriminative model, and incorporates the correlation graph into the learning procedure to capture the intrinsic manifold structures of multi-modal data. DBRC develops a deep network based on a special Multimodal Restricted Boltzmann Machine (MRBM) to learn binary codes. The network employs the adaptive tanh hash function to obtain the binary valued representation instead of joint real value representation, and reconstructs the original data to preserve the maximum semantic similarity across different modalities. DCMH first extracts the deep features of text and image modalities through two neural networks, and then preserves the similarity of two different deep features into a unified hash codes by using a pair-wise similarity matrices. DD-CMH proposes a cross-modal deep neural networks to directly encode the binary hash codes by employing discrete optimization, which can effectively preserve the intra- and inter-modality semantic correlation. SSAH devices a deep self-supervised adversarial network to solve cross-modal hashing problem. This network combines multi-label semantic information and adversarial learning to eliminate the semantic gap between deep features extracted from heterogeneous modalities.

Differences: The existing deep learning based cross-modal hashing approaches equally handle the different cross-modal retrieval tasks when constructing the hash functions. Under such circumstance, the characteristics of cross-modal retrieval tasks are ignored during the hash learning process, and thus sub-optimal performance may be achieved accordingly. Different from them, in our paper, we put forward a task-adaptive cross-modal hash learning model to learn two couples of hash functions for two cross-modal sub-retrieval tasks respectively. In our model, the semantic similarity across different modalities are preserved and the representation capability of query modality is enhanced. With such learning framework, the hash codes we learned can simultaneously capture the semantic correlation of different modalities and the query semantics of the specific cross-modal retrieval task.

Table 1: The list of main notations.

Notation	Description
$\mathbf{X}$	the raw image matrix
$\mathbf{Y} \in \mathbb{R}^{n \times d_y}$	the text feature matrix
$\mathbf{F} \in \mathbb{R}^{r \times n}$	deep feature representation matrix of image
$\mathbf{G} \in \mathbb{R}^{r \times n}$	deep feature representation matrix of text
$\mathbf{P} \in \mathbb{R}^{r \times c}$	semantic projection matrix of image
$\mathbf{W} \in \mathbb{R}^{r \times c}$	semantic projection matrix of text
$\mathbf{S} \in \mathbb{R}^{n \times n}$	pair-wise semantic matrix
$\mathbf{L} \in \mathbb{R}^{c \times n}$	point-wise semantic label
$\mathbf{B}_t \in \mathbb{R}^{r \times n}$	binary hash codes
N_t	mini-batch size
d_y	the dimension of text
c	the number of classes
r	hash code length
T	iteration numbers
t	the number of retrieval tasks

3. Task-adaptive asymmetric deep cross-modal hashing

3.1. Notations and problem definition

Assume that a database with n training instances is denoted as $\mathbf{O} = \{\mathbf{o}_i\}_{i=1}^{n}$, the training instance $\mathbf{o}_i$ is comprised of two modalities: image and text. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^{\top}$ denotes the raw image matrix. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^{\top} \in \mathbb{R}^{n \times d_y}$ represents the text feature matrix with d_y dimensions. Each of instance $\mathbf{y}_i$ is associated with instance $\mathbf{x}_i$. Besides, the point-wise semantic label is given as $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n] \in \mathbb{R}^{c \times n}$, where c is the total number of categories and $l_{ki} = 1$ implies that $\mathbf{o}_i$ belongs to class k, otherwise $l_{ki} = 0$. We define the pairwise semantic matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$, each element of which is represented as $S_{ij} \in \{0, 1\}$. When $S_{ij} = 1$, the image $\mathbf{x}_i$ is similar to the text $\mathbf{y}_i$, otherwise, when $S_{ij} = 0$, the image $\mathbf{x}_i$ is dissimilar to the text $\mathbf{y}_i$. In general, cross-modal retrieval problem (includes two modalities image and text) has two sub-retrieval tasks: one is the task of image searches text (I2T), and the other task is text searches image (T2I). The goal of our method is to learn two kinds of nonlinear hash functions $h(x)$ and $h(y)$ for different cross-modal retrieval task, where r is the length of hash codes, the binary hash codes $\mathbf{B}_1$ relates to images hash functions $h(x)$ for I2T task, and the binary hash codes $\mathbf{B}_2$ relates to texts hash functions $h(y)$ for T2I task. Table 1 shows the list of main notations used in this paper.

3.2. Model formulation

In this paper, we propose an supervised asymmetric deep cross-modal hashing model, which includes two parts: deep feature learning and asymmetric hash learning. In the first part, we extract the deep image and text feature representations from two couples of deep neural networks. In the second part, we perform asymmetric hash learning to capture the semantic correlations of multimedia data with the supervision of pair-wise semantic matrix and enhance the discriminative ability of query modality representation with point-wise semantic label. The overall learning framework of our TA-ADCMH method is illustrated in Figure 1.

3.2.1. Deep feature learning

In the deep feature learning part, we design two couples of deep neural networks for two cross-modal sub-retrieval tasks. As shown in the Figure 1, we can find that each pair of image-text deep networks are used to perform I2T and T2I sub-retrieval tasks, respectively. To be fair, we use similar deep neural networks of image modality for two sub-retrieval tasks. Both two deep networks are based on convolutional neural network (CNN-F) and use the pre-trained ImageNet dataset [36] to initialize the weights of networks. Particularly, CNN-F is an eight layer deep network structure with five layers convolution layer and three fully-connected layers. We modify the structure of last fully-connected layer by setting the length of hash codes as the number of hidden units, and adopt identity function as the activation function for the last network layers. We also use two deep neural networks of text modality for two sub-retrieval tasks, each of which consist of two fully-connected layers. Particularly, we represent the original text

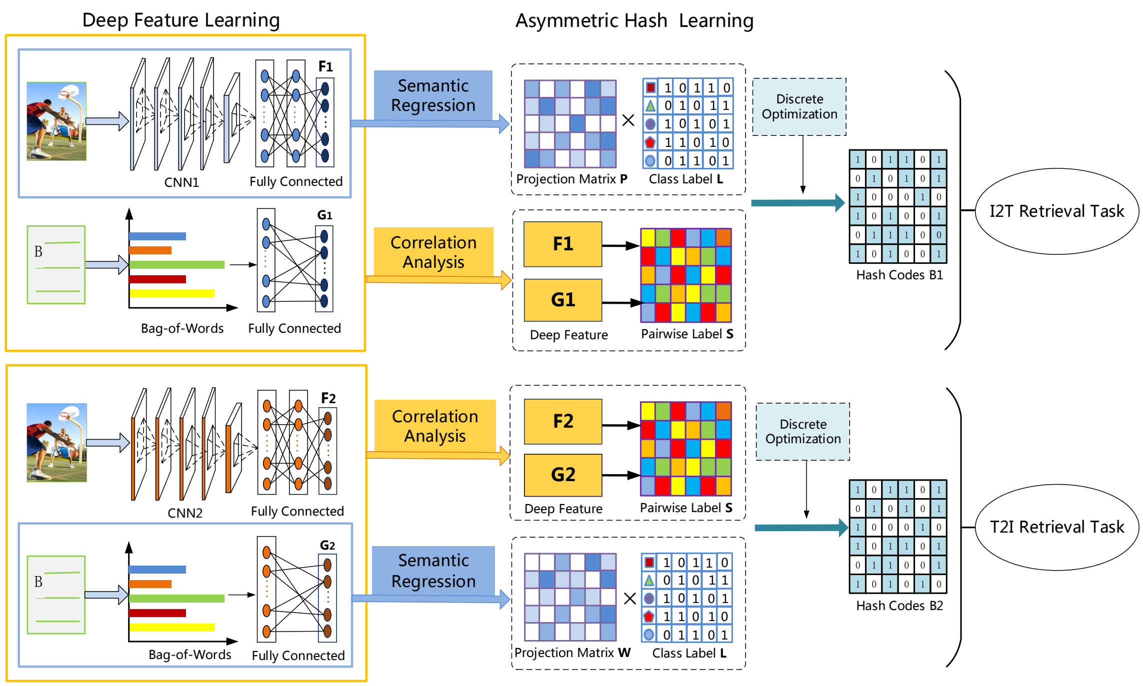


Figure 1: The overall learning framework of our TA-ADCMH method.

vectors as the Bag-of-Words (BOW) [37] which is then used as the input to the deep neural network. Further, we obtain the hash codes as the outputs from the last fully-connected layer. Similar to the image network, we also adopt identity function as the activation function. In this paper, the deep hash functions are denoted as $f_t(\mathbf{X}_i;\theta_{x_t})|_{t=1}^{2}$ for image modality and $g_t(\mathbf{Y}_j;\theta_{y_t})|_{t=1}^{2}$ text modalities separately, where $\theta_{x_t}|_{t=1}^{2}$ is the weight parameters of deep image neural networks and $\theta_{y_t}|_{t=1}^{2}$ is the weight parameters of deep text neural networks.

3.2.2. Asymmetric hash learning for I2T

The cross-modal retrieval task concentrates on two sub-retrieval tasks: image retrieves text and text retrieves image. Previous methods generally learn the same couple of hash functions in an symmetric way for two different cross-modal retrieval tasks. They cannot effectively capture the query semantics during the non-linear multi-modal mapping process, as ignoring the characteristics of different cross-modal retrieval tasks.

To address these problems, in this paper, we develop an asymmetric hash learning model to learn different hash functions for different retrieval tasks. Specifically, for each task, besides to optimize the semantic preserving of multi-modal data into hash codes, we perform the semantic regression from query-specific modality representation to the explicit labels. With such design, the semantic correlations of multi-modal data can be preserved into the hash codes, and simultaneously, the query semantics can be captured adaptively.

The overall objective function of I2T sub-retrieval task is formulated as

$$
\begin{aligned}
\min_{\mathbf{B}_1,\theta_{x_1},\theta_{y_1},\mathbf{P}} \mathcal{J}_1 = & \underbrace{-\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij} - \log(1+e^{\Phi_{ij}})) + \lambda_1\|\mathbf{B}_1-\mathbf{F}_1\|_F^2 + \beta_1\|\mathbf{B}_1-\mathbf{G}_1\|_F^2}_{\text{Correlation Analysis}} \\
& + \underbrace{\mu_1\|\mathbf{F}_1-\mathbf{P}\mathbf{L}\|_F^2 + \nu_1 R(\mathbf{F}_1\mathbf{1},\mathbf{G}_1\mathbf{1},\mathbf{P})}_{\text{Semantic Regression}} \\
& s.t.\, \mathbf{B}_1 \in \{-1,1\}^{r\times n}
\end{aligned}
\tag{1}
$$

where λ_1, β_1, μ_1, ν_1 are all the regularization parameters, $\Phi_{ij} = \frac{1}{2}\mathbf{F}_{1i}^{\top}\mathbf{G}_{1j}$, $\mathbf{F}_1 \in \mathbb{R}^{r\times n}$ with $f_1(\mathbf{X}_i;\theta_{x_1})$, $\mathbf{G}_1 \in \mathbb{R}^{r\times n}$ with $g_1(\mathbf{Y}_j;\theta_{y_1})$ are the deep features extracted from images and texts respectively. $\mathbf{B}_1 \in \mathbb{R}^{r\times n}$ is the binary hash codes to be learned for I2T task. It is binary values by imposing the discrete constraint. $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n]$ is the point-wise semantic label. $\mathbf{P} \in \mathbb{R}^{r\times c}$ is the semantic projection matrix which supports the semantic regression from image (query) modality representation $\mathbf{F}_1$ to the $\mathbf{L}$. The first term in Eq.(1) is negative log likelihood function, which is based on the likelihood function defined as

$$
p(S_{ij}|\mathbf{F}_{1i},\, \mathbf{G}_{1j}) = \begin{cases} \sigma(\Phi_{ij}) & S_{ij}=1 \\ 1-\sigma(\Phi_{ij}) & S_{ij}=0 \end{cases} \tag{2}
$$

where $\sigma(\Phi_{ij}) = \frac{1}{1+e^{-\Phi_{ij}}}$. The negative log likelihood function can make $\mathbf{F}_{1i}$ and $\mathbf{G}_{1j}$ as similar as possible when $S_{ij} = 1$, and be dissimilar when $S_{ij} = 0$. Thus, this term can preserve the semantic correlation between deep image feature $\mathbf{F}_{1i}$ and deep text feature $\mathbf{G}_{1j}$ by the pair-wise semantic supervision. The second and third terms in Eq.(1) transform the deep features $\mathbf{F}_{1i}$ and $\mathbf{G}_{1j}$ into the binary hash codes $\mathbf{B}_1$, which collectively preserve the cross-modal semantics into the binary hash codes. The last term is to avoid overfitting. It is defined as below:

$$\nu_1 R(\mathbf{F}_1\mathbf{1},\ \mathbf{G}_1\mathbf{1},\ \mathbf{P}) = \nu_1\, \|\mathbf{F}_1\mathbf{1}\|_F^2 + \nu_1\, \|\mathbf{G}_1\mathbf{1}\|_F^2 + \nu_1\, \|\mathbf{P}\|_F^2 \tag{3}$$

The term $\nu_1\, \|\mathbf{F}_1\mathbf{1}\|_F^2 + \nu_1\, \|\mathbf{G}_1\mathbf{1}\|_F^2$ ($\mathbf{1} = [1, ..., 1]^{\top} \in \mathbb{R}^n$) is to equally partition the information of each bit and ensure the maximum semantic similarity preserved into hash codes.

3.2.3. Asymmetric hash learning for T2I

Different from I2T sub-retrieval task, we directly regress the deep text representation to the corresponding point-wise semantic label to persist the discriminative information of query modality representation. Specifically, we adopt pair-wise semantic label to learn a new binary hash codes $\mathbf{B}_2$ to preserve the semantic correlation of multimodal data and capture the query semantics from texts.

Similar to Eq.(1), the objective function for T2I sub-retrieval task is formulated as:

$$\begin{aligned} \min_{\mathbf{B}_2,\theta_{x_2},\theta_{y_2},\mathbf{W}} \mathcal{J}_2 = & \underbrace{-\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij}\ -\ \log(1+e^{\Phi_{ij}})) + \lambda_2\|\mathbf{B}_2-\mathbf{F}_2\|_F^2 + \beta_2\|\mathbf{B}_2-\mathbf{G}_2\|_F^2}_{\text{Correlation Analysis}} \\ & + \underbrace{\mu_2\|\mathbf{G}_2-\mathbf{WL}\|_F^2 + \nu_2 R(\mathbf{F}_2\mathbf{1},\mathbf{G}_2\mathbf{1},\mathbf{W})}_{\text{Semantic Regression}} \\ & s.t.\ B_2 \in \{-1,1\}^{r\times n} \end{aligned} \tag{4}$$

where $\Phi_{ij} = \frac{1}{2}\mathbf{F}_{2i}^{\top}\mathbf{G}_{2j}$, $\mathbf{F}_2 \in \mathbb{R}^{r\times n}$ with $f_2(\mathbf{X}_2;\theta_{x_2})$, $\mathbf{G}_2 \in \mathbb{R}^{r\times n}$ with $g_2(\mathbf{Y}_2;\theta_{y_2})$ are the deep features extracted from images and texts respectively. $\mathbf{W} \in \mathbb{R}^{r\times c}$ is the semantic projection matrix which supports the semantic regression from text (query) modality representation $\mathbf{G}_2$ to the $\mathbf{L}$. The balance parameters $\lambda_2, \beta_2, \mu_2$ and ν_2 are regularization parameters of T2I task. The regularization function $R(\mathbf{F}_2\mathbf{2},\ \mathbf{G}_2\mathbf{2},\ \mathbf{W})$ is denoted as follows:

$$\nu_2 R(\mathbf{F}_2\mathbf{1},\ \mathbf{G}_2\mathbf{1},\ \mathbf{W}) = \nu_2\, \|\mathbf{F}_2\mathbf{1}\|_F^2 + \nu_2\, \|\mathbf{G}_2\mathbf{1}\|_F^2 + \nu_2\, \|\mathbf{W}\|_F^2 \tag{5}$$

This term $\nu_2 \|\mathbf{F}_2\mathbf{1}\|_F^2 + \nu_2 \|\mathbf{G}_2\mathbf{1}\|_F^2$ ($\mathbf{1} = [1, ..., 1]^{\mathrm{T}} \in \mathbb{R}^n$) is same as that for I2T task, which is used to balance each bit of hash codes.

3.3. Optimization scheme

The objective functions for I2T and T2I retrieval tasks are all non-convex with the involved variables. In this paper, we propose an iterative optimization method to learn the optimal value for I2T and T2I.

1. For the I2T sub-retrieval task, we give the following iterative optimization steps:

Step 1. Update θ_{x_1}. The problem in Eq.(1) can be rewritten as

$$\min_{\theta_{x_1}} -\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij} - \log(1+e^{\Phi_{ij}})) + \lambda_1\|\mathbf{B}_1 - \mathbf{F}_1\|_F^2 + \mu_1\|\mathbf{F}_1 - \mathbf{PL}\|_F^2 + \nu_1\|\mathbf{F}_1\mathbf{1}\|_F^2$$

The deep CNN parameter θ_{x_1} of image modality can be trained by stochastic gradient descent (SGD) [38] with the back-propagation (BP) algorithm. In each iteration, we randomly select a mini-batch samples from the database to train the network, which relieves the SGD algorithm from falling directly into the local optimal value near the initial point. Specifically, we first compute the following gradient for each instance of $\mathbf{x}_i$:

$$\frac{\partial \mathcal{J}_1}{\partial \mathbf{F}_{1i}} = \frac{1}{2}\sum_{j=1}^{n}(\sigma(\Phi_{ij})\mathbf{G}_{1j} - S_{ij}\mathbf{G}_{1j}) + 2\lambda_1(\mathbf{F}_{1i} - \mathbf{B}_{1i}) + 2\mu_1(\mathbf{F}_1 - \mathbf{PL}) + 2\nu_1\mathbf{F}_1\mathbf{1} \quad (6)$$

Then we can compute the $\frac{\partial \mathcal{J}_1}{\partial \theta_{x_1}}$ according to the BP updating rule until convergency.

Step 2. Update θ_{y_1}. The optimization problem in Eq.(1) becomes

$$\min_{\theta_{y_1}} -\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij} - \log(1+e^{\Phi_{ij}})) + \beta_1\|\mathbf{B}_1 - \mathbf{G}_1\|_F^2 + \nu_1\|\mathbf{G}_1\mathbf{1}\|_F^2$$

The deep CNN parameter θ_{y_1} of text modality is also trained by SGD and BP algorithm. Firstly, we compute the following gradient for each instance of y_j:

$$\frac{\partial \mathcal{J}_1}{\partial \mathbf{G}_{1j}} = \frac{1}{2}\sum_{i=1}^{n}(\sigma(\Phi_{ij})\mathbf{F}_{1i} - S_{ij}\mathbf{F}_{1i}) + 2\beta_1(\mathbf{G}_{1j} - \mathbf{B}_{1j}) + 2\nu_1\mathbf{G}_1\mathbf{1} \quad (7)$$

Then we can compute the $\frac{\partial \mathcal{J}_1}{\partial \theta_{y_1}}$ according to the BP updating rule until convergency.

Step 3. Update $\mathbf{B}_1$. The problem in Eq.(1) can be formulated as

$$\min_{\mathbf{B}_1} \lambda_1\|\mathbf{B}_1 - \mathbf{F}_1\|_F^2 + \beta_1\|\mathbf{B}_1 - \mathbf{G}_1\|_F^2 \quad s.t.\, \mathbf{B}_1 \in \{-1, 1\}^{r\times n} \quad (8)$$

The solution of Eq.(8) can be easily obtained by optimizing without relaxing discrete binary constraints $\mathbf{B}_1 \in \{-1, 1\}^{r\times n}$. Thus, we have

$$\mathbf{B}_1 = \mathtt{sgn}(\lambda_1 \mathbf{F}_1 + \beta_1 \mathbf{G}_1) \tag{9}$$

Step 4. Update **P**. The corresponding optimization problem can be simplified as

$$\min_{\mathbf{P}} \mu_1 ||\mathbf{F}_1 - \mathbf{PL}||_F^2 + \nu_1 ||\mathbf{P}||_F^2 \tag{10}$$

Letting the derivative of Eq.(10) with respect to **P** to be equal to zero, we obtain

$$\mathbf{P} = \mathbf{F}_1 \mathbf{L}^{\mathbb{T}} (\mathbf{L}\mathbf{L}^{\mathbb{T}} + \frac{\nu_1}{\mu_1} \mathbf{I})^{-1} \tag{11}$$

2. For the T2I sub-retrieval task, we give the details of the iterative optimization algorithm as shown below:

Step 1. Update θ_{y_2}. The problem in Eq.(4) can be reduced to

$$\begin{aligned} \min_{\theta_{y_2}} - \sum_{i,j=1}^{n} (S_{ij}\Phi_{ij} - \log(1 + e^{\Phi_{ij}})) + \beta_2 ||\mathbf{B}_2 - \mathbf{G}_2||_F^2 \\ + \mu_2 ||\mathbf{G}_2 - \mathbf{WL}||_F^2 + \nu_2 ||\mathbf{G}_2 \mathbf{1}||_F^2 \end{aligned} \tag{12}$$

The deep CNN parameter θ_{y_2} of text modality can be learned by SGD and BP algorithm. Firstly, we compute the following gradient for each instance of $\mathbf{y}_j$:

$$\begin{aligned} \frac{\partial \mathcal{J}_2}{\partial \mathbf{G}_{2j}} =& \frac{1}{2} \sum_{i=1}^{n} (\sigma(\Phi_{ij}) \mathbf{F}_{2i} - S_{ij} \mathbf{F}_{2i}) + 2\beta_2 (\mathbf{G}_{2j} - \mathbf{B}_{2j}) \\ &+ 2\mu_2 (\mathbf{G}_2 - \mathbf{WL}) + 2\nu_2 \mathbf{G}_2 \mathbf{1} \end{aligned} \tag{13}$$

Then we can compute the $\frac{\partial \mathcal{J}_2}{\partial \theta_{y_2}}$ according to the BP updating rule until convergency.

Step 2. Update θ_{x_2}. The optimization problem in Eq.(4) becomes

$$\min_{\theta_{x_2}} - \sum_{i,j=1}^{n} (S_{ij}\Phi_{ij} - \log(1 + e^{\Phi_{ij}})) + \lambda_2 ||\mathbf{B}_2 - \mathbf{F}_2||_F^2 + \nu_2 ||\mathbf{F}_2 \mathbf{1}||_F^2$$

The deep CNN parameter θ_{x_2} of image modality can be trained by SGD and BP algorithm. Firstly, we compute the following gradient for each instance of $\mathbf{x}_i$:

$$\frac{\partial \mathcal{J}_2}{\partial \mathbf{F}_{2i}} = \frac{1}{2} \sum_{j=1}^{n} (\sigma(\Phi_{ij}) \mathbf{G}_{2j} - S_{ij} \mathbf{G}_{2j}) + 2\lambda_2 (\mathbf{F}_{2i} - \mathbf{B}_{2i}) + 2\nu_2 \mathbf{F}_2 \mathbf{1} \tag{14}$$

Then we can compute the $\frac{\partial \mathcal{J}_{\in}}{\partial \theta_{x_2}}$ according to the BP updating rule until convergency.

Step 3. Update $\mathbf{B}_2$. The problem in Eq.(4) is rewritten as follows

$$\min_{\mathbf{B}_2} \lambda_2 \|\mathbf{B}_2 - \mathbf{F}_2\|_F^2 + \beta_2 \|\mathbf{B}_2 - \mathbf{G}_2\|_F^2 \quad s.t.\, \mathbf{B}_2 \in \{-1, 1\}^{r \times n} \tag{15}$$

Without relaxing discrete constrains, we can obtain the hash codes of Eq.(15) as

$$\mathbf{B}_2 = \mathtt{sgn}(\lambda_2 \mathbf{F}_2 + \beta_2 \mathbf{G}_2) \tag{16}$$

Step 4. Update $\mathbf{W}$. The optimization problem of Eq.(4) can be reformulated as follows

$$\min_{\mathbf{W}} \mu_2 \|\mathbf{G}_2 - \mathbf{W}\mathbf{L}\|_F^2 + \nu_2 \|\mathbf{W}\|_F^2 \tag{17}$$

The solution can also be obtained by setting the derivative of Eq.(17) with respect to $\mathbf{W}$ to be equal to zero, we obtain

$$\mathbf{W} = \mathbf{G}_2 \mathbf{L}^{\top} (\mathbf{L}\mathbf{L}^{\top} + \frac{\nu_2}{\mu_2} \mathbf{I})^{-1} \tag{18}$$

The final results can be obtained by repeating the above steps until convergence. Algorithm 1 summarizes the key optimization steps for the I2T task in the proposed TA-ADCMH.

3.4. Online query hashing

As we discussed earlier, TA-ADCMH is a deep asymmetric cross-modal hashing method. It learns task-adaptive hash functions for different retrieval tasks. Specifically, given a new query instance $\mathbf{x}_q$ with image modality, we can obtain its hash codes for I2T retrieval task by using the following formula

$$b_q = h(\mathbf{x}_q) = \mathtt{sgn}(f_1(\mathbf{x}_q; \theta_{\mathbf{x}_1})).$$

Similarly, given a query instance with text modality $\mathbf{y}_q$, we can obtain the corresponding hash codes for T2I retrieval task by

$$b_q = h(\mathbf{y}_q) = \mathtt{sgn}(g_2(\mathbf{y}_q; \theta_{\mathbf{y}_2})).$$

Algorithm 1 Discrete optimization for I2T

Input: The raw image matrix $\mathbf{X}$, text feature matrix $\mathbf{Y}$, pair-wise semantic matrix $\mathbf{S}$, point-wise semantic label $\mathbf{L}$, hash code length r, the parameters $\lambda_1, \beta_1, \mu_1, \nu_1$.

Output: Hash codes matrix $\mathbf{B}_1$, deep network parameters θ_{x_1} and θ_{y_1}.

Randomly initialize $\mathbf{P}$, $\mathbf{B}_1$, θ_{x_1}, θ_{y_1}.

Construct the mini-batch N_1 and N_2 from $\mathbf{X}$ and $\mathbf{Y}$ by randomly sampling, $N_1 = N_2$=128. Initialize the iteration number $T_1 = \lceil n/N_1 \rceil$, $T_2 = \lceil n/N_2 \rceil$

repeat

  For $iter = 1, \cdots, T_1$ **do**

    Calculate $f_1(\mathbf{X}_i; \theta_{x_1})$ by the forward propagation according to Eq.(6)

    Update deep model parameters θ_{x_1} by using back propagation.

  end for

  For $iter = 1, \cdots, T_2$ **do**

    Calculate $g_1(\mathbf{Y}_j; \theta_{y_1})$ by the forward propagation according to Eq.(7)

    Update deep model parameters θ_{y_1} by using back propagation.

  end for

    Update hash codes $\mathbf{B}_1$ according to Eq.(9).

    Update semantic projection matrix $\mathbf{P}$ according to Eq.(11).

until convergence

4. Experimental setting

4.1. Evaluation datasets

We conduct experiments on two public cross-modal retrieval datasets: MIR Flickr [39] and NUS-WIDE [40]. Both of them includes image and text modalities.

MIR Flickr includes 25,000 pairs of image-text instances collected from Flickr website. This dataset provides 24 labels and uses them to classify the instances, each of which belongs to at least one category. We select 20,015 instances labeled with no less than 20 textual tags to compose the final dataset. For the convenience, the query set of 2,000 multi-modal data were chosen by random selection, the retrieval set is composed of the remaining 18,015 multi-modal data. Within the retrieval set, the training set of 10,000 instances is further chosen on a random basis. We describe each text as a 1,386-dimensional BOW vector. To be fair, the input of shallow methods are 4,096-dimensional CNN feature, and the input of deep methods are original image pixies.

NUS-WIDE includes 269,648 instances with 81 semantic labels downloaded from Flickr website. Considering the imbalance of label distribution, we select the top 21 most common categories and ultimately obtain 195,834 image-text pairs as our final dataset. In our experiments, we choose 2,000 pairs instances for query, 193,834 pairs instances for retrieval, 10,000 pairs instances for training. The text of each instance is expressed as a 1,000-dimensional BOW vector. For the traditional methods, the image of each instances is described by a deep feature with 4,096-dimension. For the deep methods, each image uses the original pixel directly as the input.

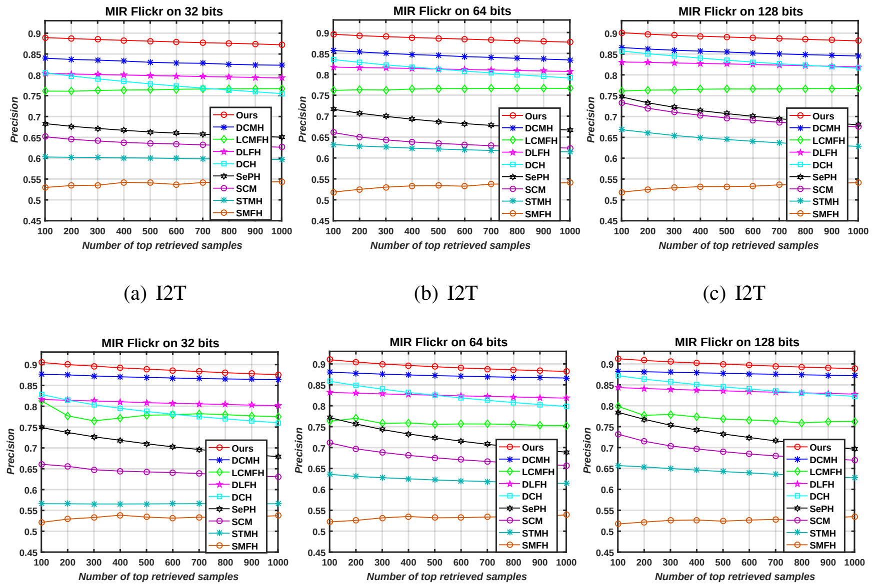
4.2. Evaluation baselines and metrics

We compare our proposed TA-ADCMH with several typical cross-modal retrieval methods, including SCM [23], SePH [24], SMFH [15], STMH [11], DCH [25], DLFH [33], LCMFH [13] and DCMH [28]. Note that there are two versions of several methods that have different optimization algorithms. In our experiment, the sequential learning is used to learn SCM method, k-means is used for SePH method, and kernel logistic regression is used for DLFH method. Among the eight compared baselines, DCMH is a deep method and the others are shallow methods. SCM and SMFH are two relaxation methods, which discard discrete constraints during the process of hash code quantization. Other methods directly adopt discrete optimization to solve the hash codes.

To evaluate the retrieval performance, we adopt mean Average Precision (mAP) [41] and topK-precision [42, 43] as the evaluation metrics.

4.3. Implementation details

Our method formulates two objective functions, which consist of eight parameters: $\lambda_1, \beta_1, \mu_1, \nu_1, \lambda_2, \beta_2, \mu_2, \nu_2$ for two different cross-modal retrieval tasks. In the task of I2T retrieval, the regularization parameters λ_1, β_1 control the regression the deep features to the asymmetric binary codes for images and text respectively. The parameter

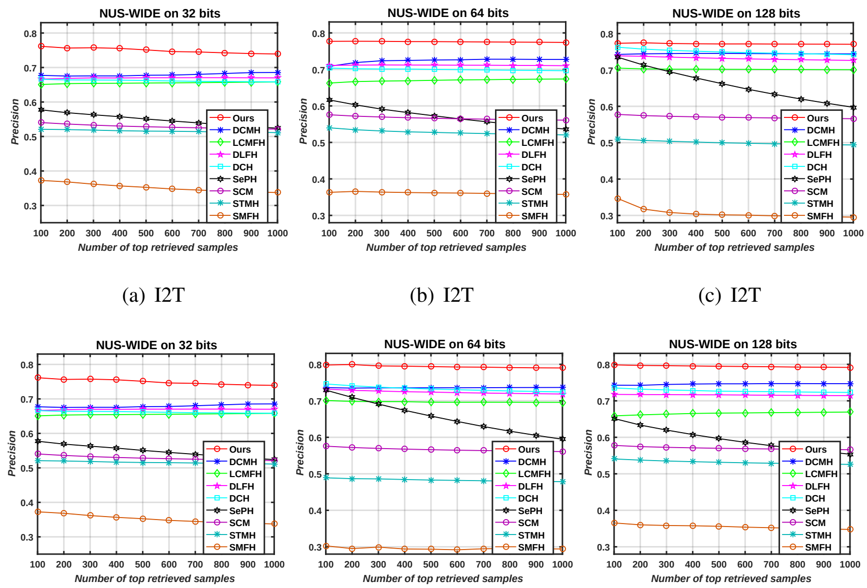


(a) I2T (b) I2T (c) I2T

(d) T2I (e) T2I (f) T2I

Figure 2: The performance of topK-precision curves on MIR Flickr.

μ_1 ensures the deep features of image modality are ideal for supervised classification, ν_1 is the regularization parameter to avoid overfitting. The parameters of the T2I retrieval task are similar to I2T task. We set different values for the involved parameters to optimize the I2T and T2I retrieval tasks. In our experiment, the best performance of I2T task can be achieved when $\{\lambda_1 = 10^{-1}, \beta_1 = 10^{-2}, \mu_1 = 10^{-4}, \nu_1 = 10^{-1}\}$, and that of T2I task can be achieved when $\{\lambda_2 = 10^{-1}, \beta_2 = 10^{2}, \mu_2 = 10^{-1}, \nu_2 = 10^{-1}\}$ on MIR Flickr. Besides, the best performance can be obtained for I2T task when $\{\lambda_1 = 10^{-1}, \beta_1 = 1, \mu_1 = 1, \nu_1 = 10^{-1}\}$, and that for T2I task when $\{\lambda_2 = 10^{-1}, \beta_2 = 1, \mu_2 = 1, \nu_2 = 10^{-1}\}$ on NUS-WIDE. In all cases, the iterations number is set to 500. Moreover, we implement our method on Matconvnet and use CNN deep networks which are the same as DCMH. Before the training process, we initializing the weight after pretreatment of the original data on the ImageNet dataset. In the network learning process, we use the raw pixels for images and the BOW vectors for texts as inputs to the deep networks, respectively. The learning rate is in the range of $[10^{-6}, 10^{-1}]$. The batch size is set to 128 for two couples of deep networks. All experiments are car-



(a) I2T (b) I2T (c) I2T

(d) I2T (e) T2I (f) T2I

Figure 3: The performance of topK-precision curves on NUS-WIDE.

ried out on a computer with Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz and 64-bit Ubuntu 16.04.6 LTS operating system.

Table 2: Retrieval performance comparison (mAP) on MIR Flickr.

Methods	I2T				T2I			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
SMFH	53.62	55.36	55.70	56.58	52.07	53.28	54.44	54.52
STMH	60.08	62.80	64.58	66.59	59.67	60.69	63.53	66.60
SCM	67.47	68.50	70.91	73.65	68.50	70.93	74.16	75.01
SePH	70.15	71.33	73.53	74.28	74.31	74.95	75.42	76.75
DLFH	81.29	82.35	83.37	83.91	84.26	85.62	86.06	87.12
DCH	80.43	81.92	82.86	83.82	81.04	83.23	83.70	84.93
LCMFH	74.32	75.29	76.86	77.18	76.47	77.52	78.79	79.21
DCMH	83.42	84.09	85.55	86.83	86.96	87.71	87.96	88.01
TA-ADCMH	**88.52**	**89.47**	**90.79**	**91.30**	**91.36**	**92.91**	**93.30**	**93.87**

5. Experimental results

5.1. Retrieval accuracy comparison

In the experiments, we first report the mAP values of eight compared methods on two datasets. The mAP results of all baselines with the hash code length ranging from 16 bits to 128 bits are presented in Table 2 and Table 3. On the basis of these results, we can reach the following conclusion: 1) For both I2T and T2I sub-retrieval tasks, TA-ADCMH is consistently outperforms all the compared methods with different codes lengths. These results clearly prove the feasibility of our method. The main reasons for the superior performance are: Firstly, TA-ADCMH trains two couples of deep neural networks to perform different retrieval tasks independently, which can enhance the nonlinear representation of deep features and capture the query semantics of two sub-retrieval tasks. Secondly, we jointly adopt the pair-wise and point-wise semantic labels to generate binary codes which can express semantic similarity of different modalities. 2) It is noteworthy that the mAP values of TA-ADCMH is in the upward trend with the code length increasing. These results demonstrate that longer binary codes have stronger discriminative capability with effective discrete optimization in our method.

Table 3: Retrieval performance comparison (mAP) on NUS-WIDE.

Methods	I2T				T2I			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
SMFH	38.12	40.90	41.18	41.34	34.03	36.70	38.53	40.52
STMH	52.58	53.68	53.76	54.75	47.92	49.56	52.06	53.08
SCM	56.33	57.83	59.08	60.33	57.97	59.43	59.50	60.81
SePH	66.51	67.71	70.46	71.44	70.44	72.03	73.78	74.54
DLFH	67.31	68.52	72.78	73.24	71.70	72.14	74.78	75.92
DCH	67.23	68.97	72.57	74.85	72.52	74.13	76.84	77.01
LCMFH	67.07	68.06	69.31	69.43	70.64	71.07	72.99	73.87
DCMH	69.83	71.86	74.07	75.42	70.42	72.26	75.08	75.64
TA-ADCMH	**75.52**	**78.92**	**79.13**	**79.39**	**79.68**	**79.13**	**80.47**	**81.32**

3) The results of most methods can have higher mAP values on the T2I task than that obtained on I2T task. This depends on the fact that text features can better reflect the semantic information of instances. 4) The deep TA-ADCMH method makes significant improvement compared with the shallow methods on two datasets. Note that, DCMH is based on deep model can have the second best performance. This phenomenon is attributed to the deep feature representation extracted by the nonlinear projections and the semantic information used in binary hash mapping. The results prove that the representation capability of deep neural network is better.

Next, we illustrate the performance of topK-precision curves from 32 bits to 128 bits on two datasets. Figure 2 plots on the MIR Flickr, and figure3 plots the performance of topK-precision on NUS-WIDE. As these figures show, TA-ADCMH can obtain the higher precision and reliability compared with other baselines on both I2T and T2I sub-retrieval tasks with different code lengths. Moreover, we can also observe that the topK-precision curves of TA-ADCMH is relatively stable with the increase of retrieved samples K. These observations are enough to prove that TA-ADCMH has strong ability to retrieve the relevant samples effectively. Compared with I2T retrieval task, the topK-precision of TA-ADCMH can obtain much better performance than the

Table 4: Performance comparison for three variants on MIR Flickr.

Methods	I2T				T2I			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
TA-ADCMH-I	88.28	88.76	89.57	91.03	90.45	91.31	91.45	92.08
TA-ADCMH-II	83.12	84.63	85.51	86.47	86.27	87.01	87.84	89.25
TA-ADCMH-III	88.52	89.47	89.75	90.77	89.68	91.48	91.76	92.03
TA-ADCMH	**88.52**	**89.47**	**90.79**	**91.30**	**91.36**	**92.91**	**93.30**	**93.87**

Table 5: Performance comparison for three variants on NUS-WIDE.

Methods	I2T				T2I			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
TA-ADCMH-I	75.46	76.58	78.11	79.17	79.43	79.97	80.86	81.16
TA-ADCMH-II	68.36	69.01	72.35	74.28	68.79	71.07	73.18	74.56
TA-ADCMH-III	69.36	75.24	76.90	77.47	78.52	79.12	80.43	81.13
TA-ADCMH	**75.52**	**78.92**	**79.13**	**79.39**	**79.68**	**80.01**	**80.47**	**81.32**

baseline methods on T2I cross-modal retrieval task. It is consistent with the mAP values in Table 2 and 3. During the practical retrieval process, users browse the website according to the ranking of retrieval results, so they are interested on the top-ranked instances in the retrieved list. Thus, TA-ADCMH is significantly outperforms the comparative methods on two sub-retrieval tasks.

To summarize, TA-ADCMH achieves superior performance on MIR Flickr and NUS-WIDE. These phenomenon validates that capturing the query semantics of different cross-modal retrieval tasks is effective when learning the cross-modal hash codes. All the results confirm the effectiveness of our designed loss functions and optimization scheme.

5.2. Effects of task-adaptive hash function learning

Our method is designed to learn task-adaptive hash functions by additionally regressing the query modality representation to the class label. With the further semantic supervision, the query-specific modality representation can effectively capture the query semantics of different cross-modal retrieval tasks. To verify the effects of this part, we design two variant methods TA-ADCMH-I and TA-ADCMH-II for performance comparison. 1) TA-ADCMH-I directly performs semantic regression from class label to the shared hash codes instead of the query-specific modality representation. Mathematically, the optimization objective function of TA-ADCMH-I becomes

$$\begin{aligned}\min_{\mathbf{B},\theta_x,\theta_y,\mathbf{V}} & -\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij} - \log(1+e^{\Phi_{ij}})) + \lambda\|\mathbf{B}-\mathbf{F}\|_F^2 + \beta\|\mathbf{B}-\mathbf{G}\|_F^2 \\ & + \mu\|\mathbf{B}-\mathbf{VL}\|_F^2 + \nu(\|\mathbf{F1}\|_F^2 + \|\mathbf{G1}\|_F^2 + \|\mathbf{V}\|_F^2) \\ & s.t. \mathbf{B} \in \{-1,1\}^{r\times n}\end{aligned}$$

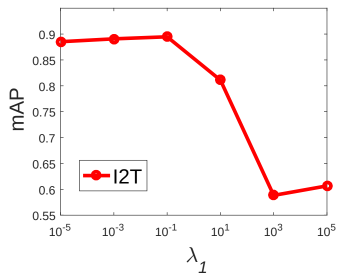
The binary codes is calculated as $\mathbf{B} = \mathtt{sgn}(\lambda\mathbf{F} + \beta\mathbf{G} + \mu\mathbf{VL})$, where λ, β, μ and ν are balance parameters. $\mathbf{L}$ is the class label. $\mathbf{V}$ is the projection matrix which regresses the hash codes $\mathbf{B}$ to the semantic label $\mathbf{L}$. The term $\nu_1 \|\mathbf{F}_1\mathbf{1}\|_F^2 + \nu_1 \|\mathbf{G}_1\mathbf{1}\|_F^2$ ($\mathbf{1} = [1, ..., 1]^{\mathrm{T}} \in \mathbb{R}^n$) is employed to make the balance of each bit of hash codes for all the training points. 2) The other variant method TA-ADCMH-II performs the pair-wise semantic supervision without employing any semantic information. Mathematically, the optimization objective function of TA-ADCMH-II becomes

$$\begin{aligned}\min_{\mathbf{B},\theta_{z_1},\theta_{z_2}} & -\sum_{i,j=1}^{n}(S_{ij}\Phi_{ij} - \log(1+e^{\Phi_{ij}})) + \lambda\|\mathbf{B}-\mathbf{Z}_1\|_F^2 + \beta\|\mathbf{B}-\mathbf{Z}_2\|_F^2 + \nu(\|\mathbf{Z}_1\mathbf{1}\|_F^2 + \|\mathbf{Z}_2\mathbf{1}\|_F^2) \\ & s.t. \mathbf{B} \in \{-1,1\}^{r\times n}\end{aligned}$$

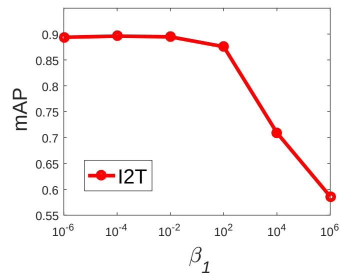
The binary codes is computed as $\mathbf{B} = \mathtt{sgn}(\lambda\mathbf{Z}_1 + \beta\mathbf{Z}_2)$, where λ, β, ν and are balance parameters. $\mathbf{Z}_1$ is the deep features of images. $\mathbf{Z}_2$ is the deep features of texts. The performance comparison results of the two variant methods are shown in Tables 4 and 5 on both two sub-retrieval tasks. Two tables demonstrate a fact that our method can outperform the variants TA-ADCMH-I and TA-ADCMH-II on two datasets with all code lengths for cross-modal retrieval. These results prove that the task-adaptive hash function learning is effective on improving the cross-modal retrieval performance.

5.3. Effects of discrete optimization

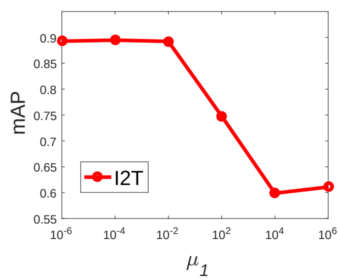
We devise a variant method TA-ADCMH-III for comparison to validate the effect of discrete optimization. Specifically, we utilize a relaxing strategy which first adopt continuous constraints instead of discrete constraints and then binarize the real-valued



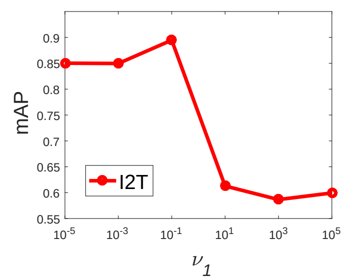
(a) The parameter of λ_1



(b) The parameter of β_1



(c) The parameter of μ_1



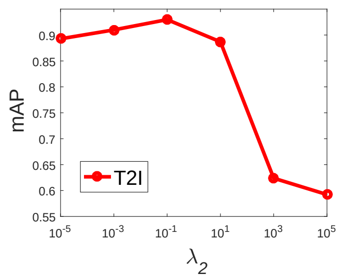
(d) The parameter of ν_1

Figure 4: Parameter experiments on I2T retrieval task.

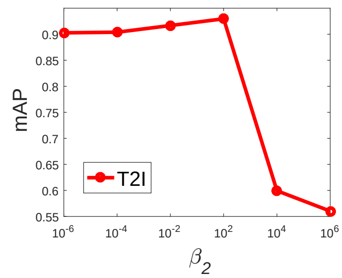
solution into hash codes by thresholding. In Eq.(1) and Eq.(4), we directly discard the constraints $\mathbf{B}_1 \in \{-1, 1\}^{r\times n}$ and $\mathbf{B}_2 \in \{-1, 1\}^{r\times n}$. Therefore, the relaxed hash codes can be calculated as $\mathbf{B}_1 = (\lambda_1\mathbf{F}_1+\beta_1\mathbf{G}_1)/(\lambda_1+\beta_1)$ for I2T task, $\mathbf{B}_2 = (\lambda_2\mathbf{F}_2+\beta_2\mathbf{G}_2)/(\lambda_2+\beta_2)$ for T2I task. Tables 4 and 5 show the comparison results of TA-ADCMH-III and TA-ADCMH on MIR Flickr and NUS-WIDE, respectively. The results demonstrate that TA-ADCMH can achieve superior performance than TA-ADCMH-III, which further validates the quantization errors are minimized by the effect discrete optimization.

5.4. Parameter experiments

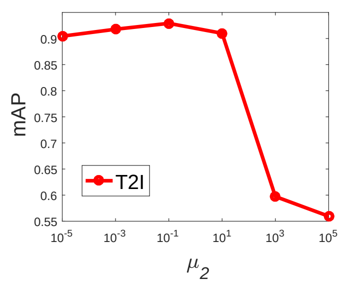
The empirical analysis of parameter sensitivity are conducted on MIR Flickr with 32 bits. Specifically, we report the mAP performance with eight involved parameters for two different retrieval tasks. In the I2T retrieval task, there are four involved parameters: $\lambda_1, \beta_1, \mu_1$ and ν_1. For λ_1 and ν_1, we tune them from $[10^{-5}, 10^{-3}, 10^{-1}, 10^{1}, 10^{3}, 10^{5}]$ by fixing the other parameters. In Figure 4 (a) and (d), the performance of λ_1 and ν_1



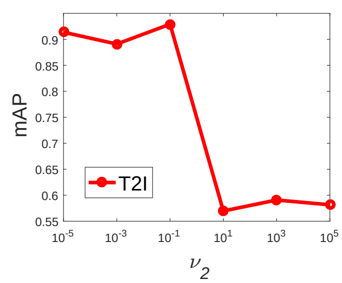
(a) The parameter of λ_2



(b) The parameter of β_2



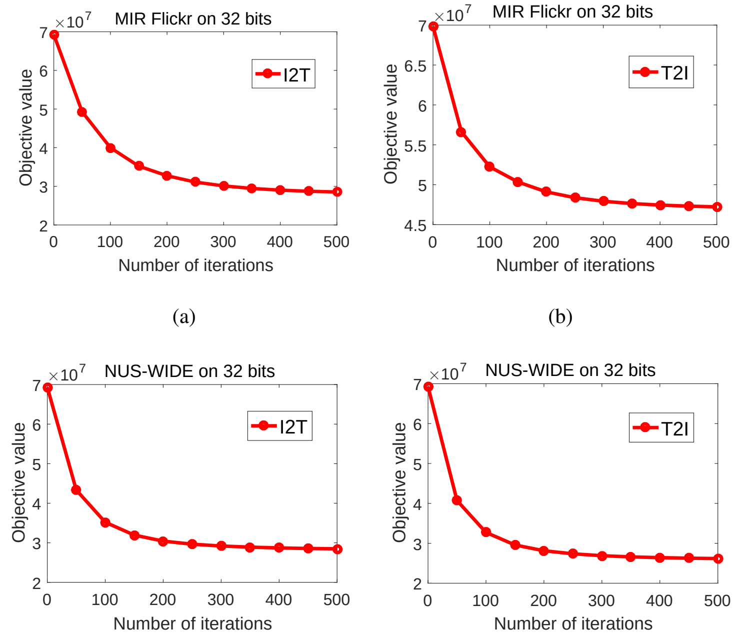
(c) The parameter of μ_2



(d) The parameter of ν_2

Figure 5: Parameter experiments on T2I retrieval task .

is relatively stable with the value of $[10^{-5}, 10^{-1}]$ for I2T retrieval task. The mAP performance decreases sharply as the two parameters are varied from 10^{-1} to 10^{5}. For the parameter β_1 and μ_1, we tune them from 10^{-6} to 10^{6}. As shown in Figure 4 (b) and (c), the I2T retrieval task is indeed influenced by β_1 and μ_1. Specifically, with β_1 increasing ($\beta_1 < 10^2$), the mAP performance is relatively stable. However, when $\beta_1 > 10^2$, the performance degrades quickly. Hence, β_1 can be chosen within the range of $[10^{-6}, 10^{-2}]$. Moreover, it is capable of observation from Figure 4 (c) that the performance of μ_1 is also relatively stable with the range of $[10^{-6}, 10^{-2}]$. The mAP values of μ_1 decreases quickly when μ_1 is larger than 10^{-2}. In particular, μ_1 can be chosen from the range between $[10^{-6}, 10^{-2}]$. The remain parameters $\lambda_2, \beta_2, \mu_2$ and ν_2 are involved for T2I retrieval task. For λ_2, μ_2, ν_2, we tune them from $[10^{-5}, 10^{-3}, 10^{-1}, 10^{1}, 10^{3}, 10^{5}]$ by fixing the other parameters. For β_2, we tune them from 10^{-6} to 10^{6}. The detailed explanation of results are shown in Figure 5. From Figure 5 (a), (c), (d), we have a conclusion that



(a) (b)

(c) (d)

Fig 6: Convergence curves on MIR Flickr and NUS-WIDE.

the mAP results of T2I task is in a steady trend when λ_2 ranges from $[10^{-5}$ to $10^1]$, μ_2 ranges from $[10^{-5}$ to $10^1]$ and ν_2 ranges from $[10^{-5}$ to $10^{-1}]$. From Figure 5 (b), it can be discovered that the best performance of T2I task can be obtained when β_2 is in $[10^{-6}, 10^2]$, and its performance decreases when β_2 is larger than 10^2. In general, we can reach the conclusion that the parameters of TA-ADCMH are of vital importance to our experiments and they can be stable within a reasonable range of values.

5.5. Convergence analysis

To analyze the convergence of TA-ADCMH, we display the convergence curves in Figure 6 with 32 bits on MIR Flickr and NUS-WIDE. For I2T retrieval task, the convergence analysis results of the formula Eq.(1) are recorded in Figure 6 (a) and (c). For T2I retrieval task, the convergence analysis results of the formula Eq.(4) are recorded in Figure 6 (b) and (d). In all the figures, abscissa is the iteration numbers and ordinate is the value of the objective function. As shown in the figures, we can find a fact that TA-ADCMH achieves a stable minimum within 300 iterations for I2T task and within 400 for T2I task on MIR Flickr dataset. We can also find that it converges within 300 iterations for both two retrieval tasks on NUS-WIDE dataset. The experimental results confirm that TA-ADCMH can converge gradually.

6. Conclusion

In this work, we propose a Task-adaptive Asymmetric Deep Cross-modal Hashing (TA-ADCMH) method. It learns task-adaptive hash functions for different cross-modal retrieval tasks. The deep learning framework jointly optimizes the semantic preserving from multi-modal deep representations to the hash codes, and the semantic regression from the query-specific representation to the explicit labels. The hash codes we learned can effectively preserve the multi-modal semantic correlations, and meanwhile, adaptively capture the query semantics. Further, we devise a discrete optimization scheme to effectively solve the discrete binary constraints of binary codes. On two datasets, we prove the superiority of our TA-ADCMH method.

References

References

[1] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H. T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the ACM International Conference on Multimedia (MM), 2017, pp. 154–162.

[2] L. Wang, L. Zhu, X. Dong, L. Liu, J. Sun, H. Zhang, Joint feature selection and graph regularization for modality-dependent cross-modal retrieval, Journal of Visual Communication and Image Representation 54 (2018) 213–222.

[3] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, M. Cord, Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings, in: Proceedings of the ACM International Conference on Research on Development in Information Retrieval (SIGIR), 2018, pp. 35–44.

[4] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, A. G. Hauptmann, Adaptive semi-supervised feature selection for cross-modal retrieval, IEEE Transactions on Multimedia 21 (5) (2018) 1276–1288.

[5] L. Wang, L. Zhu, E. Yu, J. Sun, H. Zhang, Task-dependent and query-dependent subspace learning for cross-modal retrieval, IEEE Access 6 (2018) 27091–27102.

[6] Y. Song, M. Soleymani, Polysemous visual-semantic embedding for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1979–1988.

[7] F. Shang, H. Zhang, L. Zhu, J. Sun, Adversarial cross-modal retrieval based on dictionary learning, Neurocomputing 355 (2019) 93–104.

[8] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2075–2082.

[9] T. Wang, L. Zhu, Z. Cheng, J. Li, Z. Gao, Unsupervised deep cross-modal hashing with virtual label regression, Neurocomputing (2019). doi:https://doi.org/10.1016/j.neucom.2019.12.058.

[10] L. Wang, L. Zhu, E. Yu, J. Sun, H. Zhang, Fusion-supervised deep cross-modal hashing, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 37–42.

[11] D. Wang, X. Gao, X. Wang, L. He, Semantic topic multimodal hashing for cross-media retrieval, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2015, pp. 3890–3896.

[12] L. Xie, L. Zhu, G. Chen, Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval, Multimedia Tools and Applications 75 (15) (2016) 9185–9204.

[13] D. Wang, X.-B. Gao, X. Wang, L. He, Label consistent matrix factorization hashing for large-scale cross-modal similarity search, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (10) (2018) 2466–2479.

[14] L. Xie, L. Zhu, P. Pan, Y. Lu, Cross-modal self-taught hashing for large-scale image retrieval, Signal Processing 124 (2016) 81–92.

[15] J. Tang, K. Wang, L. Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Transactions on Image Processing 25 (7) (2016) 3157–3166.

[16] L. Xie, J. Shen, L. Zhu, Online cross-modal hashing for web image retrieval, in: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2016, pp. 294–300.

[17] X. Liu, X. Nie, W. Zeng, C. Cui, L. Zhu, Y. Yin, Fast discrete cross-modal hashing with regressing from semantic labels, in: Proceedings of the ACM International Conference on Multimedia (MM), 2018, pp. 1662–1669.

[18] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2011, pp. 1360–1365.

[19] J. Song, Y. Yang, Y. Yang, Z. Huang, H. T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the ACM Conference on Management of Data (SIGMOD), 2013, pp. 785–796.

[20] X. Zhu, Z. Huang, H. T. Shen, X. Zhao, Linear cross-modal hashing for efficient multimedia search, in: Proceedings of the ACM International Conference on Multimedia (MM), 2013, pp. 143–152.

[21] V. E. Liong, J. Lu, Y.-P. Tan, Cross-modal discrete hashing, Pattern Recognition 79 (2018) 114 – 129. doi:10.1016/j.patcog.2018.02.002.

[22] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, H. T. Shen, Collective reconstructive embeddings for cross-modal hashing, IEEE Transactions on Image Processing 28 (6) (2018) 2770–2784.

[23] D. Zhang, W.-J. Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2014, pp. 2177–2183.

[24] Z. Lin, G. Ding, M. Hu, J. Wang, Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3864–3872.

[25] X. Xu, F. Shen, Y. Yang, H. T. Shen, X. Li, Learning discriminative binary codes for large-scale cross-modal retrieval, IEEE Transactions on Image Processing 26 (5) (2017) 2494–2507.

[26] J. Zhang, Y. Peng, M. Yuan, Unsupervised generative adversarial cross-modal hashing, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 539–546.

[27] D. Hu, F. Nie, X. Li, Deep binary reconstruction for cross-modal hashing, IEEE Transactions on Multimedia 21 (4) (2018) 973–985.

[28] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3232–3240.

[29] F. Zhong, Z. Chen, G. Min, Deep discrete cross-modal hashing for cross-media retrieval, Pattern Recognition 83 (2018) 64–77. `doi:10.1016/j.patcog.2018.05.018`.

[30] J. Zhou, G. Ding, Y. Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the ACM International Conference on Research on Development in Information Retrieval (SIGIR), 2014, pp. 415–424.

[31] D. Wang, Q. Wang, X. Gao, Robust and flexible discrete hashing for cross-modal similarity search, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2017) 2703–2715.

[32] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Advances in neural information processing systems(NIPS), 2009, pp. 1753–1760.

[33] Q.-Y. Jiang, W.-J. Li, Discrete latent factor model for cross-modal hashing, IEEE Transactions on Image Processing 28 (7) (2019) 3490–3501.

[34] F. Shen, C. Shen, W. Liu, H. Tao Shen, Supervised discrete hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 37–45.

[35] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4242–4251.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[37] Y. Yang, F. Shen, H. T. Shen, H. Li, X. Li, Robust discrete spectral hashing for large-scale image semantic indexing, IEEE Transactions on Big Data 1 (4) (2015) 162–171.

[38] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of the International Conference on Computational Statistics (COMPSTAT), 2010, pp. 177–186.

[39] M. J. Huiskes, M. S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the ACM International Conference on Multimedia Information Retrieval (MIR), 2008, pp. 39–43.

[40] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Viedo Retrieval (CIVR), 2009, p. 48.

[41] X. Xu, L. He, A. Shimada, R.-i. Taniguchi, H. Lu, Learning unified binary codes for cross-modal retrieval via latent semantic hashing, Neurocomputing 213 (2016) 191–203. doi:10.1016/j.neucom.2015.11.133.

[42] X. Lu, L. Zhu, Z. Cheng, X. Song, H. Zhang, Efficient discrete latent semantic hashing for scalable cross-modal retrieval, Signal Processing 154 (2019) 217 – 231. doi:10.1016/j.sigpro.2018.09.007.

[43] X.-S. Xu, Dictionary learning based hashing for cross-modal retrieval, in: Proceedings of the ACM International Conference on Multimedia (MM), 2016, pp. 177–181.