# Equally-Guided Discriminative Hashing for Cross-modal Retrieval

**Yufeng Shi**[1] , **Xinge You**[1*] , **Feng Zheng**[2] , **Shuo Wang**[1] and **Qinmu Peng**[1]

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology

[2]Department of Computer Science and Engineering, Southern University of Science and Technology

{yufengshi17, youxg}@hust.edu.cn

## Abstract

Cross-modal hashing intends to project data from two modalities into a common hamming space to perform cross-modal retrieval efficiently. Despite satisfactory performance achieved on real applications, existing methods are incapable of effectively preserving semantic structure to maintain inter-class relationship and improving discriminability to make intra-class samples aggregated simultaneously, which thus limits the higher retrieval performance. To handle this problem, we propose Equally-Guided Discriminative Hashing (EGDH), which jointly takes into consideration semantic structure and discriminability. Specifically, we discover the connection between semantic structure preserving and discriminative methods. Based on it, we directly encode multi-label annotations that act as high-level semantic features to build a common semantic structure preserving classifier. With the common classifier to guide the learning of different modal hash functions equally, hash codes of samples are intra-class aggregated and inter-class relationship preserving. Experimental results on two benchmark datasets demonstrate the superiority of EGDH compared with the state-of-the-arts.

## 1 Introduction

With the exponential growth of multimedia data in Internet, cross-modal retrieval that aims to retrieve samples from one modality with a given query from another modality, is increasingly important and draws much attention in machine learning. Despite the empirical success in information retrieval [Radenović *et al.*, 2018], cross-modal retrieval [Hwang and Grauman, 2012; Peng *et al.*, 2018] is still a challenging problem due to the heterogeneity between modalities (e.g., image and text). To better tackle this problem, hashing based methods gradually become the mainstream approach in the area in recent years due to its low memory usage and high query speed [Wang *et al.*, 2018; Peng *et al.*, 2018].

Substantial efforts have been made to remove modality heterogeneity [Ding *et al.*, 2016; Wang *et al.*, 2017; Cao *et al.*, 2018; Wu *et al.*, 2019]. One of the most well-known method is Cross-view Hashing (CVH) [Kumar and Udupa, 2011], which minimizes the similarity weighted distances between modalities. Another commonly used method, Inter-media Hashing (IMH) [Song *et al.*, 2013], encodes data to achieve the inter-modal consistency and intra-modal consistency. Moreover, Semantic Topic Multimodal Hashing (STMH) [Wang *et al.*, 2015] attempts to learn hash codes by taking into consideration latent semantic information. Despite the effective elimination of heterogeneity, aforementioned methods neglect supervised information, resulting in the performance degradation. To circumvent this problem, some work has devoted to taking fully advantage of discriminant information including semantic structure preserving based approaches [Deng *et al.*, 2018; Ma *et al.*, 2018] and discriminative approaches [Wang *et al.*, 2016; Xu *et al.*, 2016; Liu and Qi, 2018], where the former aim to express mutual similarity such as pair-wise or triplet-wise relationship, whereas the latter regard hash codes as representative features for discriminative classification.

The representative semantic structure preserving based approach, Semantic Correlation Maximization (SCM) [Zhang and Li, 2014], uses hash codes to reconstruct semantic similarity matrix. Semantics-Preserving Hashing (SePH) [Lin *et al.*, 2015] minimizes KL-divergence between the hash codes and semantics distributions. Pairwise Relationship Guided Deep Hashing (PRDH) [Yang *et al.*, 2017] was thereafter proposed to maximize pairwise semantic inter-modal similarities and intra-modal similarities. Although this type of methods significantly improve the performance by the further consideration of label information, they only focus on inter-class relationships and can not push samples of the same class to aggregate together.

As another type of approach, Multimodal Discriminative Binary Embedding (MDBE) [Wang *et al.*, 2016] formulates the hash function learning in terms of classification to obtain discriminative hash codes. Discriminant Cross-modal Hashing [Xu *et al.*, 2016] regards hash codes as features which are easily classified and builds one common classifier for different modalities. Following the same idea, Discriminative Cross-View Hashing (DCVH) [Liu and Qi, 2018] develops a neural network to fit modal-specific classifiers for differ-

[*]Contact Author

ent modalities. Compared with semantic structure preserving methods, discriminability based methods view hash codes as easily-classified features to make intra-class samples get close to each other. However, it is obvious that these discriminative methods neglect inter-class relationships which are classical focus of semantic structure preserving methods.

From the in-depth analysis, the preservation of semantic structure and the enhancement of discriminability of hash codes are equally vital for cross-modal retrieval, and it can be expected that the joint consideration can effectively improve retrieval performance. To this end, we propose Equally-Guided Discriminative Hashing (EGDH) algorithm to take into account semantic structure and discriminability in a joint manner based on the connection between classification and hashing-based retrieval tasks, resulting in discriminative hash codes which can also preserve semantic structure. Specifically, the proposed EGDH consists of three sub-networks, where the first one (i.e., labNet) is used to learn a common classifier and the last two modal-specific sub-networks (i.e., imgNet and txtNet) aim to build hash functions of different modalities. Since semantic structure preservation has been considered in the process of common classifier learning, the discriminative hash codes generated from imgNet (or txtNet) also have the ability to represent semantic structure. In conclusion, the contributions of this paper are threefold:

- We propose a novel deep cross-modal hashing method, which cooperates hashing-based retrieval with classification to preserve semantic structure and make hash codes discriminative simultaneously in a unified deep-learning framework.
- We find that angle can connect Hamming distance and classification under a special circumstance, where the intra-class samples aggregate together while preserve inter-class relationship.
- Extensive experiments on two benchmark datasets demonstrate that the proposed EGDH algorithm outperforms other baselines methods for cross-modal hashing retrieval.

The rest of this paper is organized as follows. The proposed EGDH and its optimization are introduced in Sect. 2 and experiments are conducted in Sect. 3. Finally, Sect. 4 concludes this paper.

## 2 Equally-Guided Discriminative Hashing

In this section, we present Equally-Guided Discriminative Hashing (EGDH) in detail. Here, we apply the method in two most frequently-used modalities, image and text.

### 2.1 Notation and Problem Definition

Matrix and vector used in this paper are represented as boldface uppercase letter (e.g., $\mathbf{W}$) and boldface lowercase letter (e.g., $\mathbf{w}$), respectively. $\|\cdot\|$ represents the 2-norm of vectors. $sign(\cdot)$ represents the sign function, which outputs 1 if its input is positive else outputs -1.

Let $\boldsymbol{X}^1 = \left\{\boldsymbol{x}_i^1\right\}_{i=1}^{N_1}$ and $\boldsymbol{X}^2 = \left\{\boldsymbol{x}_j^2\right\}_{j=1}^{N_2}$ represent images and texts, where $\boldsymbol{x}_i^1 \in \mathbb{R}^{d_1}$, $\boldsymbol{x}_j^2 \in \mathbb{R}^{d_2}$. And their class labels are represented by $\boldsymbol{Y} = \{y_k\}_{k=1}^{N_3}$.Following [Lin *et al.*, 2015; Cao *et al.*, 2016; Jiang and Li, 2017], we define the semantic similarity matrix $\boldsymbol{S}_{N_1 \times N_2}$ between $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$ using class labels. If $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$ share at least one category, they are similar and $S_{ij} = 1$. Otherwise, they are dissimilar and $S_{ij} = 0$. Given the code length $c$, the cross-modal hashing is to build specific hash functions $f^1\left(\boldsymbol{x}^1\right) : \mathbb{R}^{d_1} \to \{-1{,}1\}^c$ and $f^2\left(\boldsymbol{x}^2\right) : \mathbb{R}^{d_2} \to \{-1{,}1\}^c$ for images and texts. Meanwhile, the Hamming distance $D\left(\boldsymbol{h}_i^1, \boldsymbol{h}_j^2\right)$ between hash codes $\boldsymbol{h}_i^1 = f^1\left(\boldsymbol{x}_i^1\right)$ and $\boldsymbol{h}_j^2 = f^2\left(\boldsymbol{x}_j^2\right)$ indicates the similarity $S_{ij}$ between $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$. If $S_{ij} = 1$, $D\left(\boldsymbol{h}_i^1, \boldsymbol{h}_j^2\right)$ should be minimized. Otherwise, $D\left(\boldsymbol{h}_i^1, \boldsymbol{h}_j^2\right)$ should be maximized. It means:

$$S_{ij} \propto -D\left(\boldsymbol{h}_i^1, \boldsymbol{h}_j^2\right). \tag{1}$$

### 2.2 Connection

The connection between classification and hashing-based retrieval is essential in union of discriminability and semantic structure. The discriminative methods treat hash codes as features which are easily classified. Hence, classifier plays an important role. Consider a linear classifier $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{h}^T + \boldsymbol{b}$ to classify $N_3$ classes samples, where $\boldsymbol{W} = \{\boldsymbol{w}_k\}_{k=1}^{N_3}$ and $\boldsymbol{b} = \{b_k\}_{k=1}^{N_3}$. In classification tasks, the Softmax is frequently-used in converting output into normalized probability. Handled with it, the probability of $i$-th class is:

$$\begin{aligned} P_i &= \frac{\exp\left(\boldsymbol{w}_i \boldsymbol{h}_i^T + b_i\right)}{\sum_k \exp\left(\boldsymbol{w}_k \boldsymbol{h}_i^T + b_k\right)} \\ &= \frac{\exp\left(\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{h}_i\| \cdot \cos\left(\theta_{\boldsymbol{w}_i, \boldsymbol{h}_i}\right) + b_i\right)}{\sum_k \exp\left(\|\boldsymbol{w}_k\| \cdot \|\boldsymbol{h}_i\| \cdot \cos\left(\theta_{\boldsymbol{w}_k, \boldsymbol{h}_i}\right) + b_k\right)}. \end{aligned} \tag{2}$$

For Eq. (2), assume $\|\boldsymbol{w}_k\| = \|\boldsymbol{h}_i\| = \sqrt{c}$, $\boldsymbol{b}_k = 0$ and the equation is as follows:

$$\begin{aligned} P_i &= \frac{\exp\left(\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{h}_i\| \cdot \cos\left(\theta_{\boldsymbol{w}_i, \boldsymbol{h}_i}\right)\right)}{\sum_k \exp\left(\|\boldsymbol{w}_k\| \cdot \|\boldsymbol{h}_i\| \cdot \cos\left(\theta_{\boldsymbol{w}_k, \boldsymbol{h}_i}\right)\right)} \\ &= \frac{\exp\left(\sqrt{c} \cdot \sqrt{c} \cdot \cos\left(\theta_{\boldsymbol{w}_i, \boldsymbol{h}_i}\right)\right)}{\sum_k \exp\left(\sqrt{c} \cdot \sqrt{c} \cdot \cos\left(\theta_{\boldsymbol{w}_k, \boldsymbol{h}_i}\right)\right)} \\ &\propto \frac{\exp\left(\cos\left(\theta_{\boldsymbol{w}_i, \boldsymbol{h}_i}\right)\right)}{\sum_k \exp\left(\cos\left(\theta_{\boldsymbol{w}_k, \boldsymbol{h}_i}\right)\right)}. \end{aligned} \tag{3}$$

Under this circumstance, maximizing the $P_i$ equals maximizing $\cos\left(\theta_{\boldsymbol{w}_i, \boldsymbol{h}_i}\right)$ between $\boldsymbol{w}_i$ that represents the $i$-th class semantic anchor in label space and $\boldsymbol{h}_i$ that represents the feature of data in $i$-th class while minimizing $\cos\left(\theta_{\boldsymbol{w}_k, \boldsymbol{h}_i}\right)$ between $\{\boldsymbol{w}_k\}_{k \neq i}^{N_3}$ and $\boldsymbol{h}_i$. For a to-be-classified feature, the cosine between vectors in label space and itself decides its label.

Semantic structure preserving methods use hamming distance among samples to embody their relationships. When $\{-1{,}1\}^c$ composes hash codes, the norm of hash codes is the sqrt of its code length $c$, namely $\|\boldsymbol{w}_k\| = \|\boldsymbol{h}_i\| = \sqrt{c}$. The

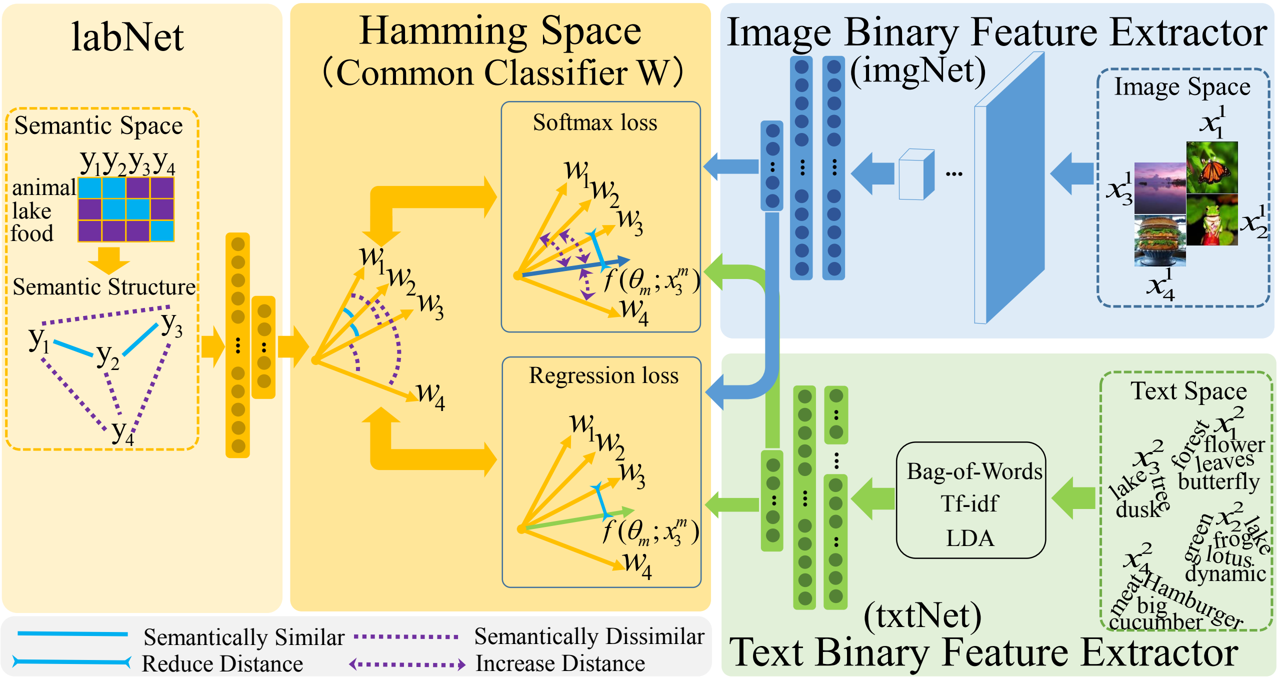

Figure 1: Equally-Guided Discriminative Hashing (EGDH) framework. Firstly, labNet encodes non-redundant multi-label annotations as hash codes to compose the common classifier $\mathbf{W}$. Later, binary feature extractor imgNet and txtNet are guided to build features which are easily classified by $\mathbf{W}$ and close to their semantic hash code anchors $\mathbf{w}$.

Hamming distance between $\boldsymbol{w}_k$ and $\boldsymbol{h}_i$ can be also measured by $\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right)$ between them:

$$\begin{aligned} D\left(\boldsymbol{w}_k,\boldsymbol{h}_i\right) &= \frac{1}{2}\left(c-\langle\boldsymbol{w}_k,\boldsymbol{h}_i\rangle\right) \\ &= \frac{1}{2}\left(c-\|\boldsymbol{w}_k\|\cdot\|\boldsymbol{h}_i\|\cdot\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right)\right) \\ &= \frac{1}{2}\left(c-\sqrt{c}\cdot\sqrt{c}\cdot\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right)\right) \\ &\propto -\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right). \end{aligned} \tag{4}$$

From Eq. (4), the Hamming distance $D\left(\boldsymbol{w}_k,\boldsymbol{h}_i\right)$ is positively correlated with $-\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right)$ of $\boldsymbol{w}_k$ and $\boldsymbol{h}_i$. Further, when combine Eq. (1) and Eq. (4), there exists a relationship:

$$S_{\boldsymbol{w}_k,\boldsymbol{h}_i} \propto \cos\left(\Theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right). \tag{5}$$

From Eq. (5), for hashing-based retrieval, the $\cos\left(\theta_{\boldsymbol{w}_k,\boldsymbol{h}_i}\right)$ indicates semantics between hash codes $\boldsymbol{w}_k$ and $\boldsymbol{h}_i$. And, the assumption that $\|\boldsymbol{w}_k\| = \sqrt{c}$, $\|\boldsymbol{h}_i\| = \sqrt{c}$ and $\boldsymbol{b}_k = 0$ is tenable in hashing retrieval when use $\{-1,1\}^c$ as hash codes and set $\boldsymbol{b} = 0$ in the classifier. Concurrently, Eq. (3) indicates that the cosine between data binary features and hash codes of label decides the label of data.

Combine Eq. (5) and Eq. (3), the cosine between vectors is a connection to integrate semantic structure and discriminability. In Hamming space, for a linear classifier composed of vectors that are already encoded to preserve their inter-class relationships, it pushes the to-be-classified binary features near corresponding semantic anchor and far away from non-corresponding anchors. By doing so, these features are the intra-class aggregated and inter-class relationship preserving. Namely, they are not only discriminative to be easily classified but also semantic structure preserving because of their cosine relationships with semantic anchors.

For cross-modal retrieval, it is also crucial to find a common space to avoid data heterogeneity. Therefore, we first build a semantic structure preserving classifier which constitutes the basics of common Hamming space. Later, the common classifier are applied to guide both binary feature extractors imgNet and txtNet to extract binary features (i.e. hash codes) which are easily classified by it and close to their semantic hash code anchors and eliminate data heterogeneity automatically.

## 2.3 Framework

As shown in Figure 1, the EGDH model consists of three parts. First, we build labNet to learn the common classifier $\mathbf{W}$. A two-layer Multi-layer Perception (MLP) whose nodes are 4096 and $c$ forms the labNet. The first layer uses ReLU as activation function and nodes in the second layer use tanh. Between these two layers, Local Response Normalization (LRN) is applied. After processing of labNet, the hash codes of non-redundant class labels are employed to compose $\mathbf{W}$. Then, we build binary feature extractors imgNet and txtNet to learn hash functions. For image, we modify CNN-F [Chatfield *et al.*, 2014] to build imgNet. To obtain $c$-length hash codes, the last fully-connected layer in origin CNN-F is changed to $c$-node fully-connected layer with tanh. For text, we first use the multi-scale network in [Li *et al.*, 2018] to extract multi-scale features and a two-layer MLP whose nodes are 4096 and $c$ to transform multi-scale features into hash codes. Except the activation function of last layer is tanh, the other layers use ReLU as activation functions. And, the LRN is also added between layers.

**Common Classifier**

For $d_3$-dimensional multi-label annotations, we regard them as high-level semantic features and transform the non-redundant multi-label annotations into $N_3$ multi-class annotations $\boldsymbol{Y} = \{\boldsymbol{y}_k\}_{k=1}^{N_3}$ $(\boldsymbol{y}_k = \{y_{k1}, y_{k2,...,} y_{kd_3}\})$. So we can use the multi-class classification method to handle multi-label classification task and traverse more non-redundant correlation. Then, the $\boldsymbol{Y} = \{\boldsymbol{y}_k\}_{k=1}^{N_3}$ are encoded as hash codes based on semantic structure. Simultaneously, these semantic hash code anchors $\{\boldsymbol{w}_1; \boldsymbol{w}_2; ...; \boldsymbol{w}_k\}$ compose the common classifier $\boldsymbol{W}$. The loss function of labNet is:

$$\begin{aligned}\min_{\boldsymbol{W},\theta_y} L^y &= L_1^y + \alpha L_2^y + \beta L_3^y \\ &= -\sum_{k,j}^{N_3} \left(S_{kj}\Delta_{kj} - \log\left(1 + e^{\Delta_{kj}}\right)\right) \\ &\quad + \alpha \sum_{k=1}^{N_3} \left(\|\boldsymbol{w}_k - f\left(\theta_y; y_k\right)\|^2\right) \\ &\quad + \beta \|\sum_{k=1}^{N_3} f\left(\theta_y; y_k\right)\|^2 \\ s.t. \boldsymbol{W} &= \{\boldsymbol{w}_k\}_{k=1}^{N_3} \in \{-1{,}1\}^c, \end{aligned} \tag{6}$$

where $\Delta_{kj} = f\left(\theta_y; \boldsymbol{y}_k\right) f\left(\theta_y; \boldsymbol{y}_j\right)^T$, $f\left(\theta_y; \boldsymbol{y}_k\right)$ is the corresponding labNet output of $\boldsymbol{y}_k$, $\boldsymbol{w}_k$ is hash codes of $f\left(\theta_y; \boldsymbol{y}_k\right)$ handled by $sign\left(\cdot\right)$, and $\alpha$, $\beta$ are hyper-parameters that adjust weights of loss. To preserve semantic structure, the first term of Eq. (6) makes hash codes to maximize likelihood of semantic similarity. As the request of hash codes, the second term aims to make the output of labNet discrete (i.e. $f\left(\theta_y; \boldsymbol{y}_k\right) = \sqrt{c}$). The last term intends to keep the number of -1 and 1 balance.

**Binary Feature Extractor**

Once the common classifier $\boldsymbol{W}$ is built, it guides binary feature extractor $f\left(\Theta_m; \boldsymbol{x}_i^m\right)$ of modality $m = 1, 2$ to obtain discriminative and semantic structure preserving hash codes equally. The loss function is as follows:

$$\begin{aligned}\min_{\theta_m} L^m &= L_1^m + \gamma L_2^m \\ &= -\sum_{i=1}^{N_m} \log\left(\frac{\exp\left(\boldsymbol{w}_i f\left(\Theta_m; \boldsymbol{x}_i^m\right)^T\right)}{\sum_k \exp\left(\boldsymbol{w}_k f\left(\Theta_m; \boldsymbol{x}_i^m\right)^T\right)}\right) \\ &\quad + \gamma \sum_{i=1}^{N_m} \left(\|\boldsymbol{w}_i - f\left(\theta_m; \boldsymbol{x}_{\boldsymbol{i}}^m\right)\|^2\right), \end{aligned} \tag{7}$$

where $\{\boldsymbol{w}_i\}_{i=1}^{N_3} \in \{-1{,}1\}^c$ are consisted of hash codes corresponding to labels and $\gamma$ is hyper-parameter. The second term is the quantization loss, which leads extractors to output binary features (i.e. $\|f\left(\theta_m; \boldsymbol{x}_{\boldsymbol{i}}^m\right)\| = \sqrt{c}$). And since the norm of elements in common classifier $\boldsymbol{w}_k$ equals $\sqrt{c}$, the connection between classification and hashing-based retrieval is practicable. Hence, the first term of Eq. (7) is the Softmax loss. Guided by the common semantic structure preserving classifier, the output features are discriminative and semantic structure preserving.

**Algorithm 1** Equally-Guided Discriminative Hashing

```
Input: Images $\boldsymbol{X}^1$, texts $\boldsymbol{X}^2$, non-redundant labels $\boldsymbol{Y}$, code length $c$, hyper-parameters $\alpha, \beta, \gamma$, learning rate $\lambda_y, \lambda_1, \lambda_2$, mini-batch size $M$, and iteration number $T_y, T_1, T_2$.
Output: Parameters $\theta_1$ and $\theta_2$ of imgNet and txtNet
  Initialization Initialize $\theta_y$, $\theta_1$, $\theta_2$ and $\boldsymbol{w}_k$.
  repeat
      for iter=1 to $T_y$ do
          Update $\theta^y$ by BP algorithm:
        $\theta_y \leftarrow \theta_y - \lambda_y \cdot \nabla_{\theta_y} L^y$
          Update $\boldsymbol{W}$ by Eq. (8)
      for iter=1 to $T_m$ do
          Update $\theta_m$ by BP algorithm:
        $\theta_m \leftarrow \theta_m - \lambda_m \cdot \nabla_{\theta_m} L^m$
  until convergence
  return $\theta_1$ and $\theta_2$
```

## 2.4 Optimization

The optimization of the EGDH model includes two parts: learning common classifier $\boldsymbol{W}$ and learning binary feature extractors $f\left(\Theta_m; \boldsymbol{x}_i^m\right)$. Learning common classifier equals to optimize $\theta_y$ and $\boldsymbol{W}$. For binary feature extractor of modality $m$, it needs to optimize $\theta_m$. The whole optimization procedure is summarized in Algorithm 1.

For $\theta_y$ of labNet, Eq. (6) is derivable. Back-propagation algorithm (BP) with mini-batch stochastic gradient descent (mini-batch SGD) method is applied to update it. As for $\boldsymbol{w}_k$, we use Eq. (8) to update it.

$$\boldsymbol{w}_k = sign\left(f\left(\theta_y; y_k\right)\right). \tag{8}$$

For binary feature extractors, we also use BP with mini-batch SGD method to update $\theta_1$ and $\theta_2$.

Once Algorithm 1 converges, the well-trained imgNet and txtNet with $sign(\cdot)$ are used to handle out-of-sample extensions from modality $m$:

$$\boldsymbol{h}_i^m = sign\left(f\left(\theta_m; \boldsymbol{x}_{\boldsymbol{i}}^m\right)\right). \tag{9}$$

# 3 Experiments

## 3.1 Datasets

Performance evaluation was conducted on two benchmark datasets: MIRFLICKR-25K [Huiskes and Lew, 2008] and MS COCO [Lin *et al.*, 2014].

MIRFLICKR-25K consists of 25015 images collected from the Flickr website. Every image is associated with several tags. Following [Jiang and Li, 2017], we use 20015 image-tag pairs annotated with 24-dimensional annotations to conduct experiments. For text data, a 1386-dimensional bag-of-words vector represents each text.

MS COCO contains 82783 training images and 40504 validation images. We eliminate images without category information and use the left 87081 images. Every image has its corresponding text description and a 91-dimensional annotation. For text, we use 2000-dimensional bag-of-words vector to represent.

## 3.2 Evaluation Protocols and Baselines

For MIRFLICKR-25K, we randomly sample 2000 image-text pairs as query set and regard the rest as retrieval set. For MS COCO, we randomly sample 5000 pairs as query set and use the rest 82081 pairs as retrieval set. For both datasets, 10000 image-text pairs are randomly chosen from retrieval set for training.

### Evaluation Protocols

To evaluate performance, we use Hamming ranking and hash lookup as retrieval protocols. Hamming ranking is to sort the data points in retrieval set based on their Hamming distance to the given query point. For comparison, we adopt mean average precision (MAP) and Top$N$-precision curves to measure it. Hash lookup aims to return retrieval data in radius of a certain Hamming distance to the given query point. We use precision-recall curve to measure its accuracy.

### Baselines

We compare EGDH with several state-of-the-art cross-modal hashing methods, including: CMSSH [Bronstein *et al.*, 2010], CVH [Kumar and Udupa, 2011], IMH [Song *et al.*, 2013], SCM [Zhang and Li, 2014], SePH [Lin *et al.*, 2015], CCQ [Long *et al.*, 2016], DCMH [Jiang and Li, 2017] and SSAH [Li *et al.*, 2018], where DCMH and SSAH are deep hashing methods. To make fair comparisons with shallow-structure-based baselines, 4096-dimensional image features extracted by the pre-trained CNN-F network are used.

## 3.3 Implementation Details

We implement all deep learning methods with Tensorflow on a NVIDIA 1080ti GPU server. Like DCMH and SSAH, we also use the CNN-F pre-trained on ImageNet [Russakovsky *et al.*, 2015] to initialize the first seven layers of imgNet. The other weights of networks are randomly initialized. We set hyper-parameters as: $\alpha = \beta = \gamma = 1$. To learn neural network parameters, we apply the Adam solver with a learning rate within $10^{-2} - 10^{-6}$ and set batch size as 128. We repeat experiments five times with random data partitions and report the averaged results.

## 3.4 Results and Discussions

In cross-modal retrieval, there are two retrieval directions: using images to query texts ($I \rightarrow T$) and using texts to query images ($T \rightarrow I$). We set the bit length at 16 bits, 32 bits, 64 bits and 128 bits.

For Hamming ranking, we report the MAP of EGDH and other baselines in Table 1. From this table, EGDH outperforms all state-of-the-art methods in longer bit lengths. Compared with the deep learning method SSAH, EGDH achieves absolute increases of 0.06%/1.72% and 2.51%/3.07% on MIRFLICKR-25K and MS COCO. EGDH achieves absolute increases on the two datasets are 3.21%/2.12% and 5.15%/2.54% compared to DCMH. Although MS COCO contains much more data than MIRFLICKR-25K, EGDH achieves more increases in MS COCO than that in MIRFLICKR-25K. It is because that hash codes obtained by EGDH are intra-class aggregated which reduces code diversity while preserve inter-class relationship. The top-$N$ curves of the two datasets are as shown in Figure 2, which also show that EGDH achieves the state-of-the-art accuracy in Hamming ranking.

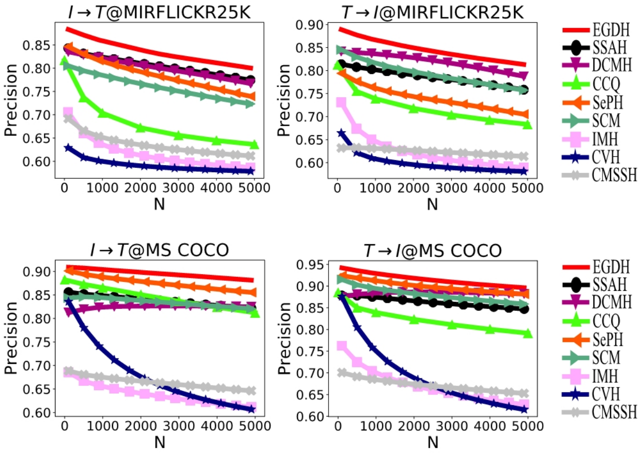

Figure 2: Top$N$-precision curves based on CNN-F feature with 128 bits code length.

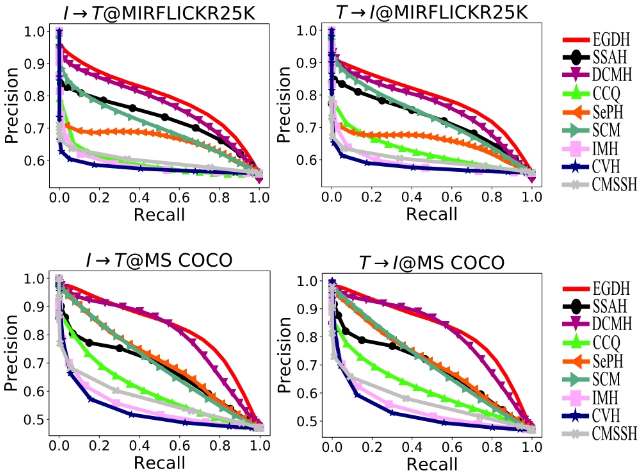

Figure 3: Precision-recall curves based on CNN-F feature with 128 bits code length.

For Hash lookup, we plot the precision-recall curves on MIRFLICKR-25K and MS COCO by varying hamming distance from 0 to 128 in Figure 3. The curves corresponding to EGDH in Figure 3 locate higher than others' on the whole. It proves that EGDH achieves the state-of-the-art efficiency in Hash lookup as like its performance in Hamming ranking.

| Task | Method | MIRFLICKR-25K | | | | MS COCO | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits |
| $I \to T$ | CMSSH [Bronstein *et al.*, 2010] | 0.6321 | 0.6201 | 0.6251 | 0.6044 | 0.5703 | 0.5674 | 0.5652 | 0.5712 |
| | CVH [Kumar and Udupa, 2011] | 0.5492 | 0.5521 | 0.5663 | 0.5754 | 0.4832 | 0.4440 | 0.4685 | 0.5236 |
| | IMH [Song *et al.*, 2013] | 0.6155 | 0.6042 | 0.5942 | 0.5852 | 0.5591 | 0.5489 | 0.5310 | 0.5428 |
| | SCM [Zhang and Li, 2014] | 0.6961 | 0.7024 | 0.7093 | 0.7125 | 0.6865 | 0.7105 | 0.7182 | 0.7240 |
| | SePH [Lin *et al.*, 2015] | 0.7213 | 0.7267 | 0.7304 | 0.7232 | 0.7568 | 0.7653 | 0.7757 | 0.7609 |
| | CCQ [Long *et al.*, 2016] | 0.6368 | 0.6282 | 0.6114 | 0.5941 | 0.6475 | 0.6348 | 0.6289 | 0.6162 |
| | DCMH [Jiang and Li, 2017] | 0.7412 | 0.7407 | 0.7504 | 0.7550 | 0.7192 | 0.7377 | 0.7477 | 0.7526 |
| | SSAH [Li *et al.*, 2018] | **0.7765** | **0.7835** | 0.7892 | 0.7643 | 0.7690 | 0.7853 | 0.7849 | 0.7238 |
| | EGDH | 0.7569 | 0.7729 | **0.7959** | **0.7900** | **0.7702** | **0.7884** | **0.8018** | **0.8029** |
| $T \to I$ | CMSSH [Bronstein *et al.*, 2010] | 0.6406 | 0.6277 | 0.6131 | 0.6047 | 0.6032 | 0.5807 | 0.5949 | 0.5745 |
| | CVH [Kumar and Udupa, 2011] | 0.5478 | 0.5511 | 0.5674 | 0.5783 | 0.4815 | 0.4411 | 0.4685 | 0.5259 |
| | IMH [Song *et al.*, 2013] | 0.6199 | 0.6083 | 0.5978 | 0.5884 | 0.5617 | 0.5525 | 0.5341 | 0.5429 |
| | SCM [Zhang and Li, 2014] | 0.7226 | 0.7340 | 0.7416 | 0.7452 | 0.6910 | 0.7168 | 0.7269 | 0.7323 |
| | SePH [Lin *et al.*, 2015] | 0.7305 | 0.7352 | 0.7364 | 0.7152 | 0.7564 | 0.7653 | 0.7764 | 0.7675 |
| | CCQ [Long *et al.*, 2016] | 0.6464 | 0.6433 | 0.6395 | 0.6296 | 0.6422 | 0.6384 | 0.6349 | 0.6248 |
| | DCMH [Jiang and Li, 2017] | 0.7629 | 0.7693 | 0.7744 | 0.7808 | 0.7475 | 0.7717 | 0.7810 | 0.7873 |
| | SSAH [Li *et al.*, 2018] | **0.7846** | 0.7935 | 0.7815 | 0.7436 | 0.7695 | 0.7891 | 0.7816 | 0.7262 |
| | EGDH | 0.7787 | **0.7939** | **0.7985** | **0.8010** | **0.7717** | **0.7935** | **0.8108** | **0.8131** |

Table 1: Mean Average Precision (MAP) comparison based on CNN-F features

| Task | Method | MIRFLICKR-25K | | | |
|---|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits | 128 bits |
| $I \to T$ | EGDH-1 | 0.7398 | 0.7502 | 0.7653 | 0.7680 |
| | EGDH-2 | 0.6417 | 0.6538 | 0.6782 | 0.6619 |
| | EGDH | **0.7569** | **0.7729** | **0.7959** | **0.7900** |
| $T \to I$ | EGDH-1 | 0.7541 | 0.7724 | 0.7876 | 0.7895 |
| | EGDH-2 | 0.6363 | 0.6559 | 0.6581 | 0.6591 |
| | EGDH | **0.7787** | **0.7939** | **0.7985** | **0.8010** |

Table 2: MAP comparison of EGDH and its variants.

### Ablation Study

For binary feature extractor, we remove the softmax loss or quantization loss while training named EGDH-1 and EGDH-2 respectively. The result is as shown in Table 2. The original EGDH outperforms EGDH-1 and EGDH-2 by 2.31%/1.71% and 12.00%/14.07% on MIRFLICKR-25K dataset. It shows the combination of discriminability and semantic structure preservation can improve the retrieval accuracy.

### Sensitivity to Parameters

We implement a parameter sensitivity experiment for the influence of hyper-parameters. Figure 4 shows the impact of $\alpha$, $\beta$ and $\gamma$ on MIRFLICKR-25K dataset at 16 bits. The EGDH is not sensitive where the hyper-parameters are within 0.01 and 2.

## 4 Conclusion

This paper introduces Equally-Guided Discriminative Hashing (EGDH) to preserve semantic structure and improve discriminability. We first build a common semantic structure preserving classifier based on the connection between classification and hashing-based retrieval. To avoid heterogeneity, the construction of hash functions for different modalities are guided by the classifier equally. Consequently, hash codes of samples are intra-class aggregated and inter-class relationship preserving. Extensive experiments show that EGDH achieves state-of-the-art results.

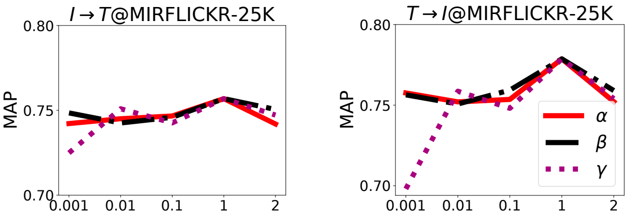

Figure 4: Influence of hyper-parameters

## Acknowledgments

This work was supported partially by the Key Science and Technology of Shenzhen (JCYJ20180305180637611), the Shenzhen Research Council (JCYJ20180305180804836 and JSGG20180507182030600), the Key Science and Technology Innovation Program of Hubei Province (2017AAA017), the Natural Science Foundation of Hubei Province (2018CFB691), the Special Projects for Technology Innovation of Hubei Province (2018ACA135), the National Natural Science Foundation of China (61571205, 61502195 and 61772220) and the Key Program for International S&T Cooperation Projects of China (2016YFE0121200).

## References

[Bronstein *et al.*, 2010] 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. IEEE, 2010.

[Cao *et al.*, 2016] Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu. Correlation autoencoder hashing for supervised cross-modal search. In *ACM ICMR*, pages 197–204. ACM, 2016.

[Cao *et al.*, 2018] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *ECCV*, pages 207–223. Springer, 2018.

[Chatfield *et al.*, 2014] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[Deng *et al.*, 2018] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *TIP*, 27(8):3893–3903, 2018.

[Ding *et al.*, 2016] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. Large-scale cross-modality search via collective matrix factorization hashing. *TIP*, 25(11):5427–5440, 2016.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ACM ICMIR*, pages 39–43. ACM, 2008.

[Hwang and Grauman, 2012] Sung Ju Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153, 2012.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *CVPR*, pages 3270–3278. IEEE, 2017.

[Kumar and Udupa, 2011] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, volume 22, page 1360, 2011.

[Li *et al.*, 2018] 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.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

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

[Liu and Qi, 2018] Liu Liu and Hairong Qi. Discriminative cross-view binary representation learning. In *WACV*, pages 1736–1744. IEEE, 2018.

[Long *et al.*, 2016] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu. Composite correlation quantization for efficient multimodal retrieval. In *ACM SIGIR*, pages 579–588. ACM, 2016.

[Ma *et al.*, 2018] Lei Ma, Hongliang Li, Fanman Meng, Qingbo Wu, and King Ngi Ngan. Global and local semantics-preserving based deep hashing for cross-modal retrieval. *NC*, 2018.

[Peng *et al.*, 2018] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *TCSVT*, 28(9):2372–2385, 2018.

[Radenović *et al.*, 2018] Filip Radenović, Giorgos Tolias, and Ondrej Chum. Fine-tuning cnn image retrieval with no human annotation. *TPAMI*, 2018.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

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

[Wang *et al.*, 2015] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. Semantic topic multimodal hashing for cross-media retrieval. In *IJCAI*, pages 3890–3896, 2015.

[Wang *et al.*, 2016] Di Wang, Xinbo Gao, Xiumei Wang, Lihuo He, and Bo Yuan. Multimodal discriminative binary embedding for large-scale cross-modal retrieval. *TIP*, 25(10):4540–4554, 2016.

[Wang *et al.*, 2017] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MC*, pages 154–162. ACM, 2017.

[Wang *et al.*, 2018] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. A survey on learning to hash. *TPAMI*, 40(4):769–790, 2018.

[Wu *et al.*, 2019] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *TIP*, 28(4):1602–1612, 2019.

[Xu *et al.*, 2016] Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Discriminant cross-modal hashing. In *ACM ICMR*, pages 305–308. ACM, 2016.

[Yang *et al.*, 2017] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, pages 1618–1625, 2017.

[Zhang and Li, 2014] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, volume 1, page 7, 2014.