This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Cross-Modal Hamming Hashing

Yue Cao, Bin Liu, Mingsheng Long(✉), and Jianmin Wang

School of Software, Tsinghua University, China
National Engineering Laboratory for Big Data Software
Beijing National Research Center for Information Science and Technology
{caoyue10,liubinthss}@gmail.com, {mingsheng,jimwang}@tsinghua.edu.cn

Abstract. Cross-modal hashing enables similarity retrieval across different content modalities, such as searching relevant images in response to text queries. It provides with the advantages of computation efficiency and retrieval quality for multimedia retrieval. Hamming space retrieval enables efficient constant-time search that returns data items within a given Hamming radius to each query, by hash lookups instead of linear scan. However, Hamming space retrieval is ineffective in existing cross-modal hashing methods, subject to their weak capability of concentrating the relevant items to be within a small Hamming ball, while worse still, the Hamming distances between hash codes from different modalities are inevitably large due to the large heterogeneity across different modalities. This work presents Cross-Modal Hamming Hashing (CMHH), a novel deep cross-modal hashing approach that generates compact and highly concentrated hash codes to enable efficient and effective Hamming space retrieval. The main idea is to penalize significantly on similar cross-modal pairs with Hamming distance larger than the Hamming radius threshold, by designing a pairwise focal loss based on the exponential distribution. Extensive experiments demonstrate that CMHH can generate highly concentrated hash codes and achieve state-of-the-art cross-modal retrieval performance for both hash lookups and linear scan scenarios on three benchmark datasets, NUS-WIDE, MIRFlickr-25K, and IAPR TC-12.

Keywords: deep hashing, cross-modal hashing, Hamming space retrieval

1 Introduction

With the explosion of big data, large-scale and high-dimensional data has been widespread in search engines and social networks. As relevant data items from different modalities may convey semantic correlations, it is significant to support cross-modal retrieval, which returns semantically-relevant results from one modality in response to a query of another modality. Recently, a popular and advantageous solution to cross-modal retrieval is learning to hash [1], an approach to approximate nearest neighbors (ANN) search across different modalities with both computation efficiency and search quality. It transforms high-dimensional data into compact binary codes with similar binary codes for similar data, largely

reducing the computational burdens of distance calculation and candidates pruning on large-scale high-dimensional data. Although the semantic gap across low-level descriptors and high-level semantics [2] has been reduced by deep learning, the intrinsic heterogeneity across modalities remains another challenge.

Previous cross-modal hashing methods capture the relations across different modalities in the process of hash function learning and transform cross-modal data into an isomorphic Hamming space, where the cross-modal distances can be directly computed [3,4,5,6,7,8,9,10,11,12,13]. Existing approaches can be roughly categorized into unsupervised methods and supervised methods. Unsupervised methods are general to different scenarios and can be trained without semantic labels or relevance information, but they are subject to the semantic gap [2] that high-level semantic labels of an object differ from low-level feature descriptors. Supervised methods can incorporate semantic labels or relevance information to mitigate the semantic gap [2], yielding more accurate and compact hash codes to improve the retrieval accuracy. However, without learning deep representations in the process of hash function learning, existing cross-modal hashing methods cannot effectively close the heterogeneity gap across different modalities.

To improve the retrieval accuracy, deep hashing methods [14,15,16] learn feature representation and hash coding more effectively using deep networks [17,18]. For cross-modal retrieval, deep cross-modal hashing methods [19,20,8,21,22,23,24] have shown that deep networks can capture nonlinear cross-modal correlations more effectively and yielded state-of-the-art cross-modal retrieval performance. Existing deep cross-modal hashing methods can be organized into unsupervised methods and supervised methods. The unsupervised deep cross-modal hashing methods adopt identical deep architecture for different modalities, e.g. MMDBM [20] uses Deep Boltzmann Machines, MSAE [8] uses Stacked Auto-Encoders, and MMNN [19] uses Multilayer Perceptrons. In contrast, the supervised deep cross-modal hashing methods [22,23,24] adopt hybrid deep architectures, which can be effectively trained with supervision to ensure best architecture for each modality, e.g. Convolutional Networks for images [17,18,25], Multilayer Perceptrons for texts [26,27,28] and Recurrent Networks for audio [29]. The supervised methods significantly outperform the unsupervised methods for cross-modal retrieval.

However, most existing methods focus on data compression instead of candidates pruning, i.e., they are designed to maximize retrieval performance by linear scan over the generated hash codes. As linear scan is still costly for large-scale database even using compact hash codes, we may deviate from our original goal towards hashing, i.e. maximizing search speedup under acceptable retrieval accuracy. With the prosperity of powerful hashing methods that perform well with linear scan, we should now return to our original ambition of hashing: enable efficient *constant-time* search using hash lookups, a.k.n. Hamming space retrieval [30]. More precisely, in Hamming space retrieval, we return data points within a given Hamming radius to each query in constant-time, by hash lookups instead of linear scan. Unfortunately, existing cross-modal hashing methods generally fall short in the capability of concentrating relevant cross-modal pairs to be within a small Hamming ball due to their mis-specified loss functions. This results in

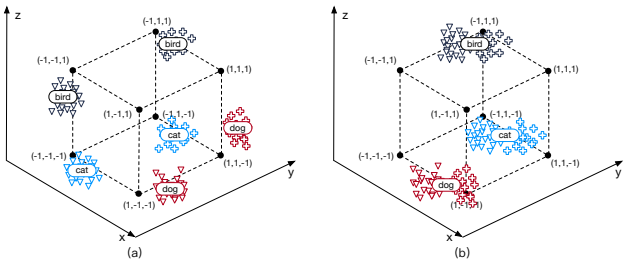


Fig. 1. Illustration of the bottleneck in cross-modal Hamming space retrieval. Different colors denote different categories (e.g. dog, cat, and bird) and different markers denote different modalities (e.g. triangles for images and crosses for texts). Due to the large intrinsic heterogeneity across different modalities, existing cross-modal hashing methods will generate hash codes of different modalities with very large Hamming distances, since their mis-specified losses cannot penalize different modalities of the same category to be similar enough in the Hamming distances, as shown in plot (a). We address this bottleneck by proposing a well-specified pairwise focal loss based on the exponential distribution, which penalizes significantly on similar cross-modal pairs with Hamming distances larger than the Hamming radius, as shown in plot (b). *Best viewed in color.*

their ineffectiveness for cross-modal Hamming space retrieval. The bottleneck of existing cross-modal hashing methods is intuitively depicted in Fig. 1.

Towards a formal solution to the aforementioned heterogeneity bottleneck in Hamming space retrieval, this work presents Cross-Modal Hamming Hashing (CMHH), a novel deep cross-modal hashing approach that generates compact and highly concentrated hash codes to enable efficient and effective Hamming space retrieval. The main idea is to penalize significantly on similar cross-modal pairs with Hamming distances larger than the Hamming radius threshold, by designing a pairwise focal loss based on the exponential distribution. CMHH simultaneously learns similarity-preserving binary representations for images and texts, and formally controls the quantization error of binarizing continuous representations to binary hash codes. Extensive experiments demonstrate that CMHH can generate highly concentrated hash codes and achieve state-of-the-art cross-modal retrieval performance for both hash lookups and linear scan scenarios on three benchmark datasets, NUS-WIDE, MIRFlickr-25K, and IAPR TC-12.

2 Related Work

Cross-modal hashing has been an increasingly important and powerful solution to multimedia retrieval [31,32,33,34,35,36]. A latest survey can be found in [1].

Previous cross-modal hashing methods include unsupervised methods and supervised methods. Unsupervised cross-modal hashing methods learn hash functions that encode data to binary codes by training from unlabeled paired data, e.g. Cross-View Hashing (CVH) [4] and Inter-Media Hashing (IMH) [7]. Supervised methods further explore the supervised information, e.g. pairwise similarity

or relevance feedbacks, to generate discriminative compact hash codes. Representative methods include Cross-Modal Similarity Sensitive Hashing (CMSSH) [3], Semantic Correlation Maximization (SCM) [11], Quantized Correlation Hashing (QCH) [12], and Semantics-Preserving Hashing (SePH) [37].

Previous shallow cross-modal hashing methods cannot exploit nonlinear correlations across different modalities to effectively bridge the intrinsic cross-modal heterogeneity. Deep multimodal embedding methods [38,39,40,41] have shown that deep networks can bridge different modalities more effectively. Recent deep hashing methods [14,15,16,42,43,44] have given state-of-the-art results on many image retrieval datasets, but they only support single-modal retrieval. There are several cross-modal deep hashing methods that use hybrid deep architectures for representation learning and hash coding, i.e. Deep Visual-Semantic Hashing (DVSH) [22], Deep Cross-Modal Hashing (DCMH) [23], and Correlation Hashing Network (CHN) [24]. DVSH is the first deep cross-modal hashing method that enables efficient image-sentence cross-modal retrieval, but it does not support the cross-modal retrieval between images and tags. DCMH and CHN are parallel works, which adopt pairwise loss functions to preserve cross-modal similarities and control quantization errors within hybrid deep architectures.

Previous deep cross-modal hashing methods fall short for Hamming space retrieval [30], i.e. hash lookups that discard irrelevant items out of the Hamming ball of a pre-specified small radius by early pruning instead of linear scan. Note that the number of hash buckets will grow exponentially with the Hamming radius and large Hamming ball will not be acceptable. The reasons for inefficient Hamming space retrieval are two folds. First, the existing methods adopt misspecified loss functions that penalize little when two similar points have large Hamming distance. Second, the huge heterogeneity across different modalities introduces large cross-modal Hamming distances. As a consequence, they cannot concentrate relevant points to be within the Hamming ball with small radius. This paper contrasts from existing methods by novel well-specified loss functions based on the exponential distribution, which shrinks the data points within small Hamming balls to enable effective hash lookups. To our best knowledge, this work is the first deep cross-modal hashing approach towards Hamming space retrieval.

3 Cross-Modal Hamming Hashing

In cross-modal retrieval, the database consists of objects from one modality and the query consists of objects from another modality. We capture the nonlinear correlation across different modalities by deep learning from a training set of N_x images $\{\boldsymbol{x}_i\}_{i=1}^{N_x}$ and N_y texts $\{\boldsymbol{y}_j\}_{j=1}^{N_y}$, where $\boldsymbol{x}_i \in \mathbb{R}^{D_x}$ denotes the D_x-dimensional feature vector of the image modality, and $\boldsymbol{y}_j \in \mathbb{R}^{D_y}$ denotes the D_y-dimensional feature vector of the text modality, respectively. Some pairs of images and texts are associated with similarity labels s_{ij}, where $s_{ij} = 1$ implies $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are similar and $s_{ij} = 0$ indicates $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are dissimilar. Deep cross-modal hashing learns modality-specific hash functions $f_x(\boldsymbol{x}) : \mathbb{R}^{D_x} \mapsto \{-1, 1\}^K$ and $f_y(\boldsymbol{y}) : \mathbb{R}^{D_y} \mapsto \{-1, 1\}^K$ through deep networks, which encode each object

$\boldsymbol{x}$ and $\boldsymbol{y}$ into compact K-bit hash codes $\boldsymbol{h}^x = f_x(\boldsymbol{x})$ and $\boldsymbol{h}^y = f_y(\boldsymbol{y})$ such that the similarity relations conveyed in the similarity pairs $\mathcal{S}$ is maximally preserved. In supervised cross-modal hashing, $\mathcal{S} = \{s_{ij}\}$ can be constructed from the semantic labels of data objects or relevance feedbacks in click-through behaviors.

Definition 1 (Hamming Space Retrieval). *For binary codes of K bits, the number of distinct hash buckets to examine is $N(K, r) = \sum_{k=0}^{r} \binom{K}{k}$, where r is the Hamming radius. $N(K, r)$ grows exponentially with r and when $r \leq 2$, it only requires $O(1)$ time for each query to find all r-neighbors. Hamming space retrieval refers to the constant-time retrieval scenario that directly returns points in the hash buckets within Hamming radius r to each query, by hash lookups.*

Definition 2 (Cross-Modal Hamming Space Retrieval). *Assuming there is an isomorphic Hamming space across different modalities, we return objects of one modality within Hamming radius r to a query of another modality, by hash lookups instead of linear scan in the modality-isomorphic Hamming space.*

This paper presents Cross-Modal Hamming Hashing (**CMHH**), a unified deep learning framework for cross-modal Hamming space retrieval, as shown in Fig. 2. The proposed deep architecture accepts pairwise inputs $\{(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij})\}$ and processes them through an end-to-end pipeline of deep representation learning and binary hash coding: **(1)** an image network to extract discriminative visual representations, and a text network to extract good text representations; **(2)** two fully-connected hashing layers for transforming the deep representations of each modality into K-bit hash codes $\boldsymbol{h}_i^x, \boldsymbol{h}_j^y \in \{1, -1\}^K$, **(3)** a new exponential focal loss based on the exponential distribution for similarity-preserving learning, which uncovers the isomorphic Hamming space to bridge different modalities, and **(4)** a new exponential quantization loss for controlling the binarization error and improving the hashing quality in the modality-isomorphic Hamming space.

3.1 Hybrid Deep Architecture

The hybrid deep architecture of CMHH is shown in Fig. 2. **For image modality**, we extend AlexNet [17], a deep convolutional network with five convolutional layers *conv1–conv5* and three fully-connected layers *fc6–fc8*. We replace the classifier layer *fc8* with a hash layer *fch* of K hidden units, which transforms the *fc7* representation into K-dimensional continuous code $\boldsymbol{z}_i^x \in \mathbb{R}^K$ for each image $\boldsymbol{x}_i$. We obtain hash code $\boldsymbol{h}_i^x$ through sign thresholding $\boldsymbol{h}_i^x = \mathrm{sgn}(\boldsymbol{z}_i^x)$. Since it is hard to optimize the sign function due to ill-posed gradient, we adopt the hyperbolic tangent (tanh) function to squash the continuous code $\boldsymbol{z}_i^x$ within $[-1, 1]$, reducing the gap between the continuous code $\boldsymbol{z}_i^x$ and the final binary hash code $\boldsymbol{h}_i^x$. **For text modality**, we follow [24,23] and adopt a two-layer Multilayer Perceptron (MLP), with the same dimension and activation function as *fc7* and *fch* in the image network. We obtain the hash code $\boldsymbol{h}_j^y$ for each text $\boldsymbol{y}_j$ also through sign thresholding $\boldsymbol{h}_j^y = \mathrm{sgn}(\boldsymbol{z}_j^y)$. To further guarantee the quality of hash codes for efficient Hamming space retrieval, we preserve the similarity between the training pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij}) : s_{ij} \in \mathcal{S}\}$ and control the quantization

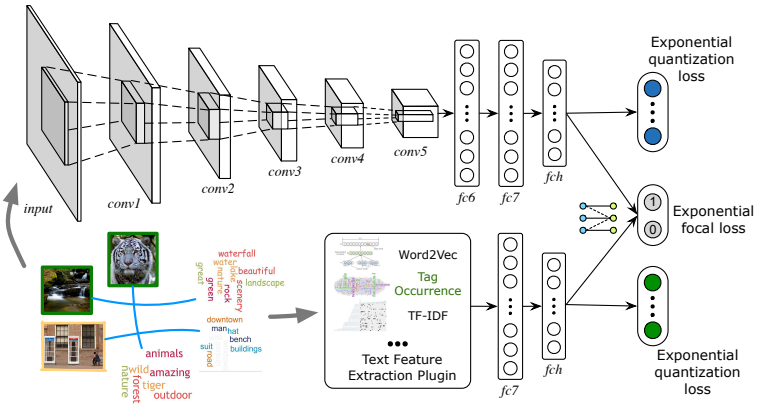


Fig. 2. The architecture of Cross-Modal Hamming Hashing (CMHH) consists of four modules: **(1)** a convolutional network for image representation and a multilayer perceptron for text representation; **(2)** two hashing layers (*fch*) for hash code generation, **(3)** an exponential focal loss for learning the isomorphic Hamming space, and **(4)** an exponential quantization loss for controlling the hashing quality. *Best viewed in color.*

error, both performed in an isomorphic Hamming space. Towards this goal, this paper proposes a pairwise exponential focal loss and a pointwise exponential quantization loss, both derived in the Maximum a Posteriori (MAP) framework.

3.2 Bayesian Learning Framework

In this paper, we propose a Bayesian learning framework to perform deep cross-modal hashing from similarity data by jointly preserving similarity relationship of image-text pairs and controlling the quantization error. Given training pairs with pairwise similarity labels as $\{(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij}) : s_{ij} \in \mathcal{S}\}$, the logarithm Maximum a Posteriori (MAP) estimation of the hash codes $\boldsymbol{H}^x = [\boldsymbol{h}_1^x, \ldots, \boldsymbol{h}_{N_x}^x]$ and $\boldsymbol{H}^y = [\boldsymbol{h}_1^y, \ldots, \boldsymbol{h}_{N_y}^y]$ for N_x training images and N_y training texts is derived as

$$\begin{aligned}\log P\left(\boldsymbol{H}^x, \boldsymbol{H}^y | \mathcal{S}\right) &\propto \log P\left(\mathcal{S} | \boldsymbol{H}^x, \boldsymbol{H}^y\right) P\left(\boldsymbol{H}^x\right) P\left(\boldsymbol{H}^y\right) \\ &= \sum_{s_{ij} \in \mathcal{S}} w_{ij} \log P\left(s_{ij} | \boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) + \sum_{i=1}^{N_x} \log P\left(\boldsymbol{h}_i^x\right) + \sum_{j=1}^{N_y} \log P\left(\boldsymbol{h}_j^y\right)\end{aligned} \tag{1}$$

where $P\left(\mathcal{S} | \boldsymbol{H}^x, \boldsymbol{H}^y\right) = \prod_{s_{ij} \in \mathcal{S}} \left[P\left(s_{ij} | \boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right]^{w_{ij}}$ is the weighted likelihood function [45], and w_{ij} is the weight for each training pair $(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij})$. For each pair, $P(s_{ij} | \boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ is the conditional probability of similarity s_{ij} given a pair of hash codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, which can be defined based on the Bernoulli distribution,

$$\begin{aligned}P\left(s_{ij} | \boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) &= \begin{cases} \sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right), & s_{ij} = 1 \\ 1 - \sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right), & s_{ij} = 0 \end{cases} \\ &= \sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)^{s_{ij}} \left(1 - \sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)^{1 - s_{ij}}\end{aligned} \tag{2}$$

where $\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)$ denotes the Hamming distance between hash codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, and σ is a probability function to be elaborated in the next subsection. Similar to binary-class logistic regression for pointwise data, we require in Equation (2) that the smaller $\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)$ is, the larger $P\left(1|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)$ will be, implying that the image-text pair $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ should be classified as similar; otherwise, the larger $P\left(0|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)$ will be, implying that the image-text pair should be classified as dissimilar. Thus, this is a natural extension of binary-class logistic regression to pairwise classification scenario with binary similarity labels $s_{ij} \in \{0, 1\}$.

Motivated by the focal loss [46], which yields state-of-the-art performance for object detection tasks, we focus our model more on hard and misclassified image-text pairs, by defining the weighting coefficient w_{ij} for each pair $(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij})$ as

$$w_{ij} = \begin{cases} \left(1 - \sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)^{\gamma}, & s_{ij} = 1 \\ \left(\sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)^{\gamma}, & s_{ij} = 0 \end{cases} \tag{3}$$

where $\gamma \geq 0$ is a hyper-parameter to control the relative weight for misclassified pairs. In Fig. 3(a), we plot the focal loss with different $\gamma \in [0, 5]$. When $\gamma = 0$, the focal loss degenerates to the standard cross-entropy loss. As γ gets larger, the focal loss gets smaller on the highly confident pairs (easy pairs), resulting in relatively more focus on the less confident pairs (hard and mis-specified pairs).

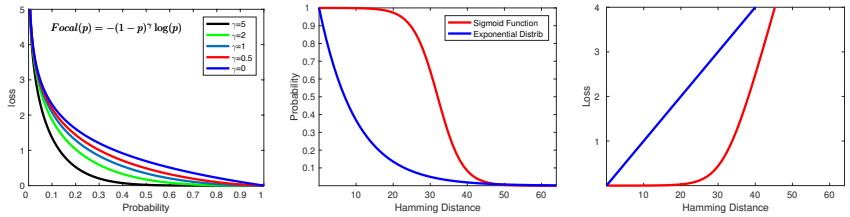
3.3 Exponential Hash Learning

With the Bayesian learning framework, any probability function σ and distance function d can be used to instantiate a specific hashing model. Previous state-of-the-art deep cross-modal hashing methods, such as DCMH [23], usually adopt the sigmoid function $\sigma(x) = 1/(1 + e^{-\alpha x})$ as the probability function, where $\alpha > 0$ is a hyper-parameter controlling the saturation zone of the sigmoid function. To comply with the sigmoid function, we need to adopt inner product as a surrogate to quantify the Hamming distance, i.e. $\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle$.

However, we discover a key *mis-specification* problem of the sigmoid function as illustrated in Fig. 3. We observe that the probability of the sigmoid function stays high when the Hamming distance between hash codes is much larger than 2 and only starts to decrease obviously when the Hamming distance becomes close to $K/2$. This implies that previous deep cross-modal hashing methods are ineffective to pull the Hamming distance between the hash codes of similar points to be smaller than 2, because the probabilities for different Hamming distances smaller than $K/2$ are not discriminative enough. This is a severe disadvantage of the existing cross-modal hashing methods, which makes hash lookup search inefficient. Note that for each query in the Hamming space retrieval, we can only return objects within the Hamming ball with a small radius (e.g. 2).

Towards the aforementioned mis-specification problem of sigmoid function, we propose a novel probability function based on the exponential distribution:

$$\sigma\left(\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right) = \exp\left(-\beta \cdot \mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right), \tag{4}$$



(a) Focal Loss (b) Probability (c) Loss

Fig. 3. **[Focal Loss]** The values of the focal loss (a) with respect to the conditional probability of similar data points ($s_{ij} = 1$). **[Exponential Distribution]** The values of Probability (b) and Loss (c) with respect to Hamming Distance between the hash codes of similar data points ($s_{ij} = 1$). The Probability (Loss) based on sigmoid function is large (small) even for Hamming distance much larger than 2, which is ill-specified for Hamming space retrieval. As a desired property, our loss based on the exponential distribution penalizes significantly on similar data pairs with larger Hamming distances.

where β is the scaling parameter of the exponential distribution, and d is the Hamming distance. In Fig. 3(b)-3(c), the probability of the exponential distribution decreases very fast when the Hamming distance gets larger than 2, and the similar points will be pulled to be within small Hamming radius. The decaying speed of the probability will be even faster by using a larger β, which imposes more force to concentrate similar points to be within small Hamming balls. Thus the scaling parameter β is crucial to control the tradeoff between precision and recall. By simply varying β, we can support a variety of Hamming space retrieval scenarios with different Hamming radiuses for different pruning ratios.

As discrete optimization of Equation (1) with binary constraints $\boldsymbol{h}_i^* \in \{-1, 1\}^K$ is challenging, continuous relaxation is applied to the binary constraints for ease of optimization, as adopted by most previous hashing methods [1,16,23]. To control the quantization error $\|\boldsymbol{h}_i^* - \mathrm{sgn}(\boldsymbol{h}_i^*)\|$ caused by continuous relaxation and to learn high-quality hash codes, we propose a novel prior distribution for each hash codes $\boldsymbol{h}_i^*$ based on a symmetric variant of the exponential distribution as

$$P\left(\boldsymbol{h}_i^*\right) = \exp\left(-\lambda \cdot \mathrm{d}\left(|\boldsymbol{h}_i^*|, \mathbf{1}\right)\right), * \in \{x, y\}, \tag{5}$$

where λ is the scaling parameter of the symmetric exponential distribution, and $\mathbf{1} \in \mathbb{R}^K$ is the vector of ones. By using the continuous relaxation, we need to replace the Hamming distance with its best approximation on continuous codes. Here we adopt Euclidean distance as the approximation of Hamming distance,

$$\mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \left\|\boldsymbol{h}_i^x - \boldsymbol{h}_j^y\right\|_2^2. \tag{6}$$

By taking Equations (2)$\sim$(5) into the MAP estimation in (1), we obtain the optimization problem of the proposed Cross-Modal Hamming Hashing (CMHH):

$$\min_{\Theta} L + \lambda Q, \tag{7}$$

where λ is a hyper-parameter to trade-off the exponential focal loss L and the exponential quantization loss Q, and Θ denotes the set of network parameters to be optimized. Specifically, the proposed *exponential focal loss* L is derived as

$$\begin{aligned} L = \sum_{s_{ij}\in\mathcal{S}} \big[& s_{ij}\left(1-\exp\left(-\beta \mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)^{\gamma} \beta \mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) \\ & - \left(1-s_{ij}\right)\left(\exp\left(-\beta \mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)^{\gamma} \log\left(1-\exp\left(-\beta \mathrm{d}\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)\right)\big], \end{aligned} \tag{8}$$

and similarly, the proposed *exponential quantization loss* is derived as

$$Q = \sum\nolimits_{i=1}^{N_x} \mathrm{d}\left(\left|\boldsymbol{h}_i^x\right|, \mathbf{1}\right) + \sum\nolimits_{j=1}^{N_y} \mathrm{d}\left(\left|\boldsymbol{h}_j^y\right|, \mathbf{1}\right), \tag{9}$$

where $\mathrm{d}(\cdot,\cdot)$ is the Hamming distance between the hash codes or the Euclidean distance between the continuous codes. Since the quantization error will be controlled by the proposed exponential quantization loss, for ease of optimization, we can use continuous relaxation for hash codes $\boldsymbol{h}_i^*$ during training. Finally, we obtain K-bit binary codes by sign thresholding $\boldsymbol{h} \leftarrow \mathrm{sgn}(\boldsymbol{h})$, where $\mathrm{sgn}(\boldsymbol{h})$ is the sign function on vectors that for $i = 1, \ldots, K$, $\mathrm{sgn}(h_i) = 1$ if $h_i > 0$, otherwise $\mathrm{sgn}(h_i) = -1$. Note that, since we have minimized the quantization error during training, the final binarization step will incur negligible loss of retrieval accuracy.

4 Experiments

We conduct extensive experiments to evaluate the efficacy of the proposed CMHH with several state-of-the-art cross-modal hashing methods on three benchmark datasets: **NUS-WIDE** [47], **MIRFlickr-25K** [48] and **IAPR TC-12** [49].

4.1 Setup

NUS-WIDE [47] is a public image dataset containing 269,648 images. Each image is annotated by some of the 81 ground truth concepts (categories). We follow similar experimental protocols as [8,50], and use the subset of 195,834 image-text pairs that belong to some of the 21 most frequent concepts.

MIRFlickr-25K [48] consists of 25,000 images coupled with complete manual annotations, where each image is labeled with some of the 38 concepts.

IAPR TC-12 [49] consists of 20,000 images with 255 concepts. We follow [23] to use the entire dataset, with each text represented as a 2912-dimensional bag-of-words vector.

We follow dataset split as [24]. In NUS-WIDE, we randomly select 100 pairs per class as the query set, 500 pairs per class as the training set and 50 pairs per class as the validation set, with the rest as the database. In MIRFlickr-25K and IAPR TC-12, we randomly select 1000 pairs as the query set, 4000 pairs as the training set and 1000 pairs as the validation set, with the rest as the database.

Following standard protocol as in [23,11,37,24], the similarity information for hash learning and for ground-truth evaluation is constructed from semantic

labels: if the image i and the text j share at least one label, they are similar and $s_{ij} = 1$; otherwise, they are dissimilar and $s_{ij} = 0$. Note that, although we use semantic labels to construct the similarity information, the proposed approach CMHH can learn hash codes when only similarity information is available.

We compare CMHH with eight state-of-the-art cross-modal hashing methods: two unsupervised methods **IMH** [7] and **CVH** [4] and six supervised methods **CMSSH** [3], **SCM** [11], **SePH** [37], **DVSH** [22], **CHN** [24] and **DCMH** [23], where **DVSH**, **CHN** and **DCMH** are deep cross-modal hashing methods.

To verify the effectiveness of the proposed CMHH approach, we first evaluate the comparison methods in the **general setting** of cross-modal retrieval widely adopted by previous methods: using linear scan instead of hash lookups. We follow [37,23,24] and adopt two evaluation metrics: Mean Average Precision (**MAP**) with MAP@R = 500, and precision-recall curves (**P@R**).

Then we evaluate **Hamming space retrieval**, following evaluation methods in [30], consisting of two consecutive steps: **(1) Pruning**, to return data points within Hamming radius 2 for each query using hash lookups; **(2) Scanning**, to re-rank the returned data points in ascending order of their distances to each query using the continuous codes. To evaluate the effectiveness of Hamming space retrieval, we report two standard evaluation metrics to measure the quality of the data points within Hamming radius 2: Precision curves within Hamming Radius 2 (**P@H≤2**), and Recall curves within Hamming Radius 2 (**R@H≤2**).

For shallow hashing methods, we use AlexNet [17] to extract 4096-dimensional deep $fc7$ features for each image. For all deep hashing methods, we directly use raw image pixels as the input. We adopt AlexNet [17] as the base architecture, and implement CMHH in **TensorFlow**. We fine-tune the ImageNet-pretrained AlexNet and train the hash layer. For the text modality, all deep methods use tag occurrence vectors as the input and adopt a two-layer Multilayer Perceptron (MLP) trained from scratch. We use mini-batch SGD with 0.9 momentum and cross-validate the learning rate from 10^{-5} to 10^{-2} with a multiplicative step-size $10^{\frac{1}{2}}$. We fix the mini-batch size as 128 and the weight decay as 0.0005. We select the hyper-parameters λ, β and γ of the proposed CMHH by cross-validation. We also select the hyper-parameters of each comparison method by cross-validation.

4.2 General Setting Results

The **MAP** results of all the comparison methods are demonstrated in Table 1, which shows that the proposed CMHH substantially outperforms all the comparison methods by large margins. Specifically, compared to SCM, the best shallow cross-modal hashing method with deep features as input, CMHH achieves absolute increases of **5.3%/7.9%**, **12.5%/19.0%** and **4.6%/8.5%** in average MAP for two cross-modal retrieval tasks $I \rightarrow T / T \rightarrow I$ on NUS-WIDE, MIRFlickr-25K, and IAPR TC-12 respectively. CMHH outperforms DCMH, the state-of-the-art deep cross-modal hashing method, by large margins of **3.5%/4.3%**, **2.9%/2.6%** and **5.0%/1.4%** in average MAP on the three benchmark datasets, respectively. Note that, compared to DVSH, the state-of-the-art deep cross-modal hashing method with well-designed architecture for image-sentence retrieval, CMHH still

Table 1. Mean Average Precision (MAP) of All Methods for Cross-Modal Retrieval.

Task	Method	NUS-WIDE			MIRFlickr-25K			IAPR TC-12		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
$I \rightarrow T$	CMSSH [3]	0.445	0.457	0.535	0.493	0.511	0.565	0.345	0.337	0.348
	CVH [4]	0.433	0.421	0.418	0.662	0.680	0.675	0.379	0.369	0.362
	IMH [7]	0.517	0.599	0.580	0.651	0.669	0.673	0.463	0.490	0.510
	SCM [11]	0.663	0.695	0.729	0.668	0.683	0.679	0.588	0.611	0.628
	SePH [37]	0.575	0.582	0.576	0.721	0.744	0.747	0.507	0.513	0.515
	DVSH [22]	-	-	-	-	-	-	0.570	0.632	0.696
	CHN [24]	0.701	0.719	0.736	0.764	0.787	0.814	0.563	0.613	0.652
	DCMH [23]	0.697	0.715	0.728	0.748	0.771	0.798	0.578	0.606	0.631
	CMHH	**0.733**	**0.738**	**0.774**	**0.783**	**0.814**	**0.821**	**0.603**	**0.657**	**0.703**
$T \rightarrow I$	CMSSH [3]	0.401	0.478	0.411	0.425	0.433	0.458	0.363	0.377	0.365
	CVH [4]	0.418	0.403	0.406	0.568	0.592	0.579	0.379	0.367	0.364
	IMH [7]	0.601	0.653	0.687	0.597	0.611	0.616	0.516	0.526	0.534
	SCM [11]	0.642	0.688	0.711	0.583	0.598	0.605	0.588	0.605	0.620
	SePH [37]	0.581	0.587	0.603	0.618	0.624	0.633	0.471	0.480	0.481
	DVSH [22]	-	-	-	-	-	-	0.604	0.640	0.681
	CHN [24]	0.671	0.712	0.736	0.719	0.748	0.761	0.647	0.683	0.695
	DCMH [23]	0.678	0.723	0.750	0.731	0.763	0.784	0.659	0.674	0.691
	CMHH	**0.719**	**0.749**	**0.778**	**0.758**	**0.782**	**0.793**	**0.667**	**0.689**	**0.710**

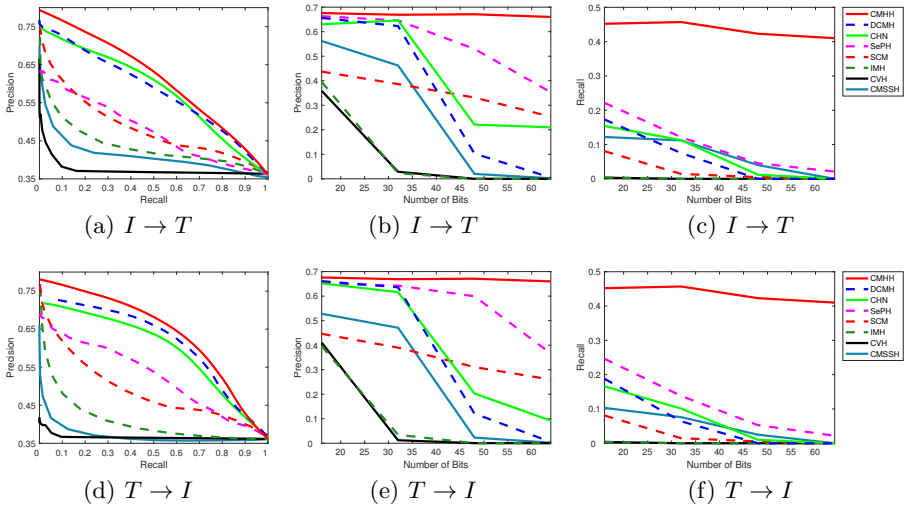
outperforms DVSH of **2.2%**/**4.7%** in average MAP for two retrieval tasks on image-sentence dataset, IAPR TC-12. This validates that CMHH is able to learn high-quality hash codes for cross-modal retrieval based on linear scan.

The proposed CMHH improves substantially from the state-of-the-art DVSH, CHN and DCMH by two key perspectives: **(1)** CMHH enhances deep learning to hash by the novel exponential focal loss motivated from the Weighted Maximum Likelihood (WML), which puts more focus on hard and misclassified examples to yield better cross-modal search performance. **(2)** CMHH learns the isomorphic Hamming space and controls the quantization error, which better approximates the cross-modal Hamming distance and learns higher-quality hash codes.

The cross-modal retrieval results in terms of Precision-Recall curves (**P@R**) on NUS-WIDE and MIRFlickr-25K are shown in Fig. 4(a), 4(d) and 5(a), 5(d), respectively. CMHH significantly outperforms all comparison methods by large margins with different lengths of hash codes. In particular, CMHH achieves much higher precision at lower recall levels or at smaller number of top returned samples. This is desirable for precision-first retrieval in practical search systems.

4.3 Hamming Space Retrieval Results

The Precision within Hamming Radius 2 (**P@H≤2**) is very crucial for Hamming space retrieval, as it only requires $O(1)$ time for each query and enables very efficient candidates pruning. As shown in Fig. 4(b), 4(e), 5(b) and 5(e), CMHH achieves the highest P@H≤2 performance on the benchmark datasets with regard to different code lengths. This validates that CMHH can learn much compacter



(a) $I \rightarrow T$ (b) $I \rightarrow T$ (c) $I \rightarrow T$

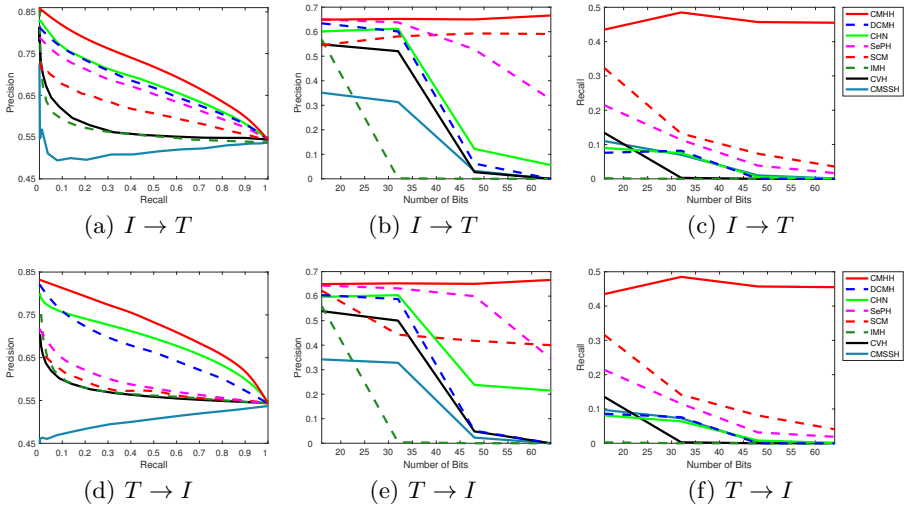
(d) $T \rightarrow I$ (e) $T \rightarrow I$ (f) $T \rightarrow I$

Fig. 4. Precision-recall (P@R) (a)(d), Precision within Hamming Radius 2 (P@H≤2) (b)(e) and Recall within Hamming Radius 2 (R@H≤2) (c)(f) on NUS-WIDE @ 32 bits.

and highly concentrated hash codes than all comparison methods and can enable more efficient and accurate Hamming space retrieval. Note that most previous hashing methods achieve worse retrieval performance with longer code lengths. This undesirable effect arises since the Hamming space will become increasingly sparse with longer code lengths and fewer data points will fall in the Hamming ball of radius 2. It is worth noting that CMHH achieves a relatively mild decrease or even an increase in accuracy using longer code lengths, validating that CMHH can concentrate hash codes of similar points together to be within Hamming radius 2, which is beneficial to Hamming space retrieval.

The Recall within Hamming Radius 2 (**R@H≤2**) is more critical in Hamming space retrieval, since it is possible that all data points will be pruned out due to the highly sparse Hamming space. As shown in Fig. 4(c), 4(f), 5(c) and 5(f), CMHH achieves the highest R@H≤2 results on both benchmark datasets with different code lengths. This validates that CMHH successfully concentrates more relevant points to be within the Hamming ball of radius 2.

It is important to note that, as the Hamming space becomes sparser using longer hash codes, most hashing baselines incur intolerable performance drop on R@H≤2, i.e. **their R@H≤2 approaches zero!** This special result reveals that existing cross-modal hashing methods cannot concentrate relevant points to be within Hamming ball with small radius, which is key to Hamming space retrieval. By introducing the novel exponential focal loss and exponential quantization loss, the proposed CMHH incurs very small performance drop on R@H≤2 as the hash codes become longer, showing that CMHH can concentrate more relevant points to be within Hamming ball with small radius even using longer code lengths. The



(a) $I \rightarrow T$ (b) $I \rightarrow T$ (c) $I \rightarrow T$

(d) $T \rightarrow I$ (e) $T \rightarrow I$ (f) $T \rightarrow I$

Fig. 5. Precision-recall (P@R) (a)(d), Precision within Hamming Radius 2 (P@H≤2) (b)(e) and Recall within Hamming Radius 2 (R@H≤2) (c)(f) on MIRFlickr @ 32 bits.

ability to adopt longer codes gives CMHH the flexibility to tradeoff accuracy and efficiency, while this is impossible for all previous cross-modal hashing methods.

4.4 Empirical Analysis

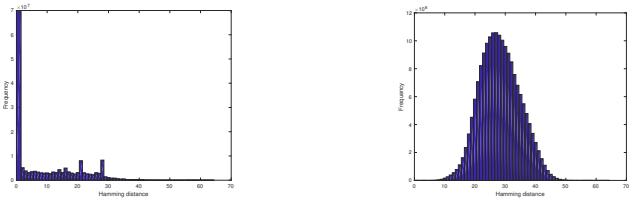
Ablation Study We investigate three variants of CMHH: **(1) CMHH-E** is the variant by replacing the exponential focal loss with the popular cross-entropy loss [23]; **(2) CMHH-F** is the variant without using the focal reweight, namely $w_{ij} = 1$ in Equation (3); **(3) CMHH-Q** is the variant without using the exponential quantization loss (9), namely λ=0; The MAP results of the three variants on the three datasets are reported in Table 2 (general setting by linear scan).

Exponential Focal Loss. (1) CMHH outperforms CMHH-E by margins of **2.7%**/**3.9%**, **2.4%**/**2.1%** and **3.6%**/**1.2%** in average MAP for cross-modal retrieval on NUS-WIDE, MIRFlickr-25K and IAPR TC-12, respectively. The exponential focal loss (8) leverages the exponential distribution to concentrate relevant points to be within small Hamming ball to enable effective cross-modal retrieval, while the sigmoid cross-entropy loss cannot achieve this desired effect. **(2)** CMHH outperforms CMHH-F by margins of **2.0%**/**2.8%**, **2.5%**/**2.1%** and **2.2%**/**2.8%** in average MAP for cross-modal tasks on the three datasets. The exponential focal loss enhances deep hashing by putting more focus on the hard and misclassified examples, and obtain better cross-modal search accuracy.

Exponential Quantization Loss. CMHH outperforms CMHH-Q by **1.9%** /**2.2%**, **1.7%**/ **2.0%** and **2.6%**/**2.3%** on the three datasets, respectively. These results validate that the exponential quantization loss (9) can boost the pruning efficiency and improve the performance of constant-time cross-modal retrieval.

Table 2. Mean Average Precision (MAP) Comparison of Different CMHH Variants.

Task	Method	NUS-WIDE			MIRFlickr-25K			IAPR TC-12		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
$I \to T$	CMHH	**0.733**	**0.738**	**0.774**	**0.783**	**0.814**	**0.821**	**0.603**	**0.657**	**0.703**
	CMHH-Q	0.708	0.715	0.765	0.762	0.788	0.804	0.578	0.623	0.685
	CMHH-F	0.710	0.721	0.753	0.755	0.779	0.798	0.589	0.631	0.677
	CMHH-E	0.705	0.722	0.736	0.751	0.780	0.802	0.584	0.619	0.653
$T \to I$	CMHH	**0.719**	**0.749**	**0.778**	**0.758**	**0.782**	**0.793**	**0.667**	**0.689**	**0.710**
	CMHH-Q	0.722	0.728	0.763	0.733	0.778	0.786	0.639	0.661	0.697
	CMHH-F	0.718	0.720	0.758	0.742	0.771	0.780	0.642	0.658	0.682
	CMHH-E	0.684	0.725	0.754	0.737	0.769	0.788	0.661	0.675	0.695



(a) CMHH (b) DCMH

Fig. 6. Histogram of Hamming distances on similar pairs @ 64 bits of CMHH & DCMH.

Statistics Study We compute the histogram of Hamming distances ($0 \sim 64$ for 64 bits codes) over all cross-modal pairs with $s_{ij} = 1$, as shown in Fig. 6. Due to the large heterogeneity across images and texts, the cross-modal Hamming distances computed based on the baseline DCMH hash codes are generally much larger than the Hamming ball radius (typically 2). This explains its nearly zero R@H$\leq$2 in Fig. 4 and 5. In contrast, the majority of the cross-modal Hamming distances computed based on our CMHH hash codes are smaller than the Hamming ball radius, which enables successful cross-modal Hamming space retrieval.

5 Conclusion

This paper establishes constant-time cross-modal Hamming space retrieval by presenting a novel Cross-Modal Hamming Hashing (CMHH) approach that can generate compacter and highly concentrated hash codes. This is done by jointly optimizing a novel exponential focal loss and an exponential quantization loss in a Bayesian learning framework. Experiments show that CMHH yields state-of-the-art cross-modal retrieval results for Hamming space retrieval and linear scan scenarios on the three datasets, NUS-WIDE, MIRFlickr-25K, and IAPR TC-12.

6 Acknowledgements

This work is supported by National Key R&D Program of China (2016YFB1000701), and National Natural Science Foundation of China (61772299, 61672313, 71690231).

References

1. Wang, J., Zhang, T., Sebe, N., Shen, H.T., et al.: A survey on learning to hash. IEEE transactions on pattern analysis and machine intelligence (2017)
2. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. TPAMI **22** (2000)
3. Bronstein, M., Bronstein, A., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: CVPR, IEEE (2010)
4. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: IJCAI. (2011)
5. Zhen, Y., Yeung, D.: Co-regularized hashing for multimodal data. In: NIPS. (2012) 1385–1393
6. Zhen, Y., Yeung, D.Y.: A probabilistic model for multimodal hash function learning. In: SIGKDD, ACM (2012)
7. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: SIGMOD, ACM (2013)
8. Wang, W., Ooi, B.C., Yang, X., Zhang, D., Zhuang, Y.: Effective multi-modal retrieval based on stacked auto-encoders. In: VLDB, ACM (2014)
9. Yu, Z., Wu, F., Yang, Y., Tian, Q., Luo, J., Zhuang, Y.: Discriminative coupled dictionary hashing for fast cross-media retrieval. In: SIGIR, ACM (2014)
10. Liu, X., He, J., Deng, C., Lang, B.: Collaborative hashing. In: CVPR, IEEE (2014)
11. Zhang, D., Li, W.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI. (2014)
12. Wu, B., Yang, Q., Zheng, W., Wang, Y., Wang, J.: Quantized correlation hashing for fast cross-modal search. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. (2015)
13. Long, M., Cao, Y., Wang, J., Yu, P.S.: Composite correlation quantization for efficient multimodal retrieval. In: SIGIR. (2016)
14. Xia, R., Pan, Y., Lai, H., Liu, C., Yan, S.: Supervised hashing for image retrieval via image representation learning. In: Proceedings of the AAAI Conference on Artificial Intellignece (AAAI), AAAI (2014)
15. Lai, H., Pan, Y., Liu, Y., Yan, S.: Simultaneous feature learning and hash coding with deep neural networks. In: CVPR. (2015)
16. Zhu, H., Long, M., Wang, J., Cao, Y.: Deep hashing network for efficient similarity retrieval. In: Proceedings of the AAAI Conference on Artificial Intellignece (AAAI), AAAI (2016)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2012)
18. Lin, M., Chen, Q., Yan, S.: Network in network. In: International Conference on Learning Representations (ICLR), 2014 (arXiv:1409.1556). (2014)
19. Masci, J., Bronstein, M.M., Bronstein, A.M., Schmidhuber, J.: Multimodal similarity-preserving hashing. IEEE Trans. Pattern Anal. Mach. Intell. **36** (2014)
20. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. JMLR **15** (2014)
21. Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J.: Deep learning for content-based image retrieval: A comprehensive study. In: MM, ACM (2014)

22. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: SIGKDD. (2016) 1445–1454
23. Jiang, Q., Li, W.: Deep cross-modal hashing. In: CVPR, 2017. (2017) 3270–3278
24. Cao, Y., Long, M., Wang, J.: Correlation hashing network for efficient cross-modal retrieval. In: BMVC. (2017)
25. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. TPAMI **35** (2013)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013)
27. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Advances in neural information processing systems. (2014)
28. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. MIT Press (1986)
29. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: ICML, ACM (2014) 1764–1772
30. Fleet, D.J., Punjani, A., Norouzi, M.: Fast search in hamming space with multi-index hashing. In: CVPR, IEEE (2012)
31. Wu, F., Yu, Z., Yang, Y., Tang, S., Zhang, Y., Zhuang, Y.: Sparse multi-modal hashing. IEEE Trans. Multimedia **16**(2) (2014) 427–439
32. Ou, M., Cui, P., Wang, F., Wang, J., Zhu, W., Yang, S.: Comparing apples to oranges: a scalable solution with heterogeneous hashing. In: SIGKDD, ACM (2013)
33. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: CVPR. (2014)
34. Wang, D., Gao, X., Wang, X., He, L.: Semantic topic multimodal hashing for cross-media retrieval. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. (2015)
35. Hu, Y., Jin, Z., Ren, H., Cai, D., He, X.: Iterative multi-view hashing for cross media indexing. In: MM, ACM (2014)
36. Wei, Y., Song, Y., Zhen, Y., Liu, B., Yang, Q.: Scalable heterogeneous translated hashing. In: SIGKDD, ACM (2014)
37. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: CVPR. (2015)
38. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015)
39. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: NIPS. (2013) 2121–2129
40. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. In: NIPS. (2014)
41. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question answering. In: NIPS. (2015)
42. Cao, Y., Long, M., Wang, J., Zhu, H., Wen, Q.: Deep quantization network for efficient image retrieval. In: Proceedings of the AAAI Conference on Artificial Intellignece (AAAI), AAAI (2016)
43. Cao, Z., Long, M., Wang, J., Yu, P.S.: Hashnet: Deep learning to hash by continuation. In: ICCV 2017. (2017)
44. Liu, B., Cao, Y., Long, M., Wang, J., Wang, J.: Deep triplet quantization. In: MM, ACM (2018)

45. Dmochowski, J.P., Sajda, P., Parra, L.C.: Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. Journal of Machine Learning Research (JMLR) **11**(Dec) (2010) 3313–3332
46. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV 2017. (2017)
47. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: CIVR, ACM (2009)
48. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: ICMR, ACM (2008)
49. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: International Workshop OntoImage. (2006) 13–23
50. Zhu, X., Huang, Z., Shen, H.T., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: MM, ACM (2013)