Correlation Hashing Network for Efficient Cross-Modal Retrieval

Yue Cao[1]
caoyue10@gmail.com
Mingsheng Long*[1]
mingsheng@tsinghua.edu.cn
Jianmin Wang[1]
jimwang@tsinghua.edu.cn
Philip Yu[12]
psyu@uic.edu

[1] KLiss, MOE; TNList; School of Software
Tsinghua University
Beijing, China

[2] University of Illinois at Chicago
Chicago, USA

Abstract

Hashing is widely applied to approximate nearest neighbor search for large-scale multimodal retrieval with storage and computation efficiency. Cross-modal hashing improves the quality of hash coding by exploiting semantic correlations across different modalities. Existing cross-modal hashing methods first transform data into low-dimensional feature vectors, and then generate binary codes by another separate quantization step. However, suboptimal hash codes may be generated since the quantization error is not explicitly minimized and the feature representation is not jointly optimized with the binary codes. This paper presents a Correlation Hashing Network (CHN) approach to cross-modal hashing, which jointly learns good data representation tailored to hash coding and formally controls the quantization error. The proposed CHN is a hybrid deep architecture that constitutes a convolutional neural network for learning good image representations, a multilayer perceptrons for learning good text representations, two hashing layers for generating compact binary codes, and a structured max-margin loss that integrates all things together to enable learning similarity-preserving and high-quality hash codes. Extensive empirical study shows that CHN yields state of the art cross-modal retrieval performance on standard benchmarks.

1 Introduction

While large-scale, high-dimensional multimedia big data are pervasive in search engines and social networks, cross-modal retrieval has attracted increasing attention, which enables approximate nearest neighbors (ANN) search across different modalities with computation efficiency and search quality. As relevant data from different modalities (image and text) may endow semantic correlations, it is important to support cross-modal retrieval that returns semantically-relevant results of one modality in response to a query of different modality. A promising solution to the cross-modal retrieval is hashing methods [28], which transform high-dimensional data into compact binary codes and generate similar binary codes for similar data. This paper focuses on cross-modal hashing that builds data-dependent hash coding

for efficient cross-media retrieval [23]. Due to large volumes and the semantic gap [24], effective cross-modal hashing remains a challenge.

Existing cross-modal hashing methods construct correlation across different modalities in the process of hash function learning and indexes cross-modal data into an isomorphic Hamming space [1, 14, 19, 20, 25, 29, 31, 33, 34, 35, 36]. They can be categorized into unsupervised methods and supervised methods. While unsupervised methods are general and can be trained without semantic labels or relevance feedbacks, they are restricted by the semantic gap [24] that high-level semantic description of an object differs from low-level feature descriptors. Supervised methods can incorporate semantic labels or relevance feedbacks to mitigate the semantic gap [24] and improve the hashing quality, i.e. achieve accurate search with shorter codes.

Recently, deep hashing methods [15, 32] have shown that both feature representation and hash coding can be learned more effectively using deep neural networks [13, 16], which can naturally encode nonlinear hashing functions. Other cross-modal retrieval models via deep learning [12, 21, 26, 27, 29] have shown that deep models can capture nonlinear cross-modal correlations more effectively and yielded state-of-the-art results on many benchmarks. However, a crucial disadvantage of these cross-modal deep hashing methods is that the quantization error is not statistically minimized hence the feature representation is not optimally compatible with binary hash coding. Another potential limitation is that they generally do not adopt principled pairwise loss function to link the pairwise Hamming distances with the pairwise similarity labels which is crucial to close the gap between the Hamming distance on binary codes and the metric distance on representations. Therefore, suboptimal representation and hash coding may be produced by existing cross-modal deep hashing methods.

This paper presents Correlation Hashing Network (CHN), a hybrid deep architecture for cross-modal hashing. CHN jointly learns good image and text representations tailored to hash coding and formally controls the quantization error, which constitutes four components: (1) an image network with multiple convolution-pooling layers to extract good image representations, and a text network with multiple fully-connected layers to extract good text representations; (2) two hashing layers to generate hash codes for each modality; (3) a cosine max-margin loss for capturing cross-modal correlation structure; and (4) a new quantization max-margin loss for controlling quality of the binarized hash codes. Extensive experiments show that CHN yields state-of-the-art results on standard cross-modal retrieval datasets.

2 Related Work

Cross-modal hashing has been a popular research topic in machine learning, computer vision, and multimedia retrieval [1, 9, 14, 19, 20, 25, 29, 33, 34, 35, 36]. We refer readers to [28] for a comprehensive survey.

Existing cross-modal hashing methods can be categorized into unsupervised methods and supervised methods. IMH [25] and CVH [14] are unsupervised methods that extend spectral hashing [30] to multimodal data. CMSSH [1], SCM [34] and QCH [31] are supervised methods, which require that if two points are known to be similar, then their corresponding hash codes from different modalities should be made similar. Since supervised methods can exploit semantic labels or relevance information to distill cross-modal correlation and reduce semantic gap [24], they can achieve superior accuracy than unsupervised methods for cross-modal similarity search with shorter hash codes.

Prior cross-modal hashing methods based on shallow architectures cannot effectively ex-

ploit the correlation across different modalities. Deep multimodal embedding methods [7] have shown that deep models can bridge heterogeneous modalities more effectively for image description. Recent deep hashing methods [3, 5, 15, 32, 37] have given state of the art results, but they can only be used for single-modal retrieval. To our knowledge, Deep Visual-Semantic Hashing (DVSH) [7] and Deep Cross-Modal Hashing (DCMH) [12] are the only two cross-modal deep hashing methods that use deep networks for representation learning and hash coding. However, our method shares the same problem setting with DCMH that only requires similarity labels across images and texts, while DVSH further requires bimodal image-text pairs to learn *modal-shared* representation. As the most similar work to ours, DCMH adopts inner product between continuous representations as the approximation to the Hamming distance between binary codes, which is not appropriate since the former takes values in $(-\infty,+\infty)$ while the latter takes values in $[-b,+b]$ (b is the number of bits). Furthermore, DCMH adopts Iterative Quantization (ITQ) [8] to generate binary codes, which may be not robust to outlier bits when the codes are unbalanced. Our CHN jointly maximizes cross-modal correlation and controls quantization error in a hybrid deep architecture with well-specified loss functions.

3 Correlation Hashing Network

In cross-modal retrieval, the database consists of objects from one modality and the query consists of objects from another modality. We uncover the correlation structure underlying different modalities by 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 vectors 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 $s_{ij}=-1$ indicates $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are dissimilar. In supervised hashing, $\mathcal{S}=\{s_{ij}\}$ can be constructed from the semantic labels of data points or the relevance feedback in click-through data. The goal of CHN is to jointly learn two modality-specific hashing functions $f_x(\boldsymbol{x}):\mathbb{R}^{d_x}\mapsto\{-1,1\}^b$ and $f_y(\boldsymbol{y}):\mathbb{R}^{d_y}\mapsto\{-1,1\}^b$ which respectively encode each unimodal point $\boldsymbol{x}$ and $\boldsymbol{y}$ in compact b-bit hash code $\boldsymbol{h}_x=f_x(\boldsymbol{x})$ and $\boldsymbol{h}_y=f_y(\boldsymbol{y})$ such that similarity information conveyed in the given bimodal object pairs $\mathcal{S}$ is maximally preserved. The Correlation Hashing Network (CHN) is a hybrid deep architecture for supervised learning to hash (Figure 1), which accepts input in a pairwise form $(\boldsymbol{x}_i,\boldsymbol{y}_j,s_{ij})$ and processes them through an end-to-end pipeline of deep representation learning and binary hash encoding.

3.1 Hybrid Deep Architecture

The hybrid deep architecture for learning cross-modal hash functions are shown in Figure 1, which constitutes an image network and a text network. In the image network, we extend AlexNet [13], a deep convolutional neural network (CNN) comprised of five convolutional layers $conv1$–$conv5$ and three fully connected layers $fc6$–$fc8$. We replace the $fc8$ layer with a new fch hash layer with b hidden units, which transforms the network activation $\boldsymbol{u}_i$ in b-bit hash code by sign thresholding $\boldsymbol{h}_i^x=\mathrm{sgn}(\boldsymbol{u}_i)$. In text network, we adopt the Multilayer perceptrons (MLP) comprising three fully connected layers, of which the last layer is replaced with a new fch hash layer with b hidden units to transform the network activation $\boldsymbol{v}_i$ in b-bit hash code by sign thresholding $\boldsymbol{h}_i^y=\mathrm{sgn}(\boldsymbol{v}_i)$. We adopt the hyperbolic tangent (tanh) function to squash the activations to be within $[-1,1]$, which reduces the gap

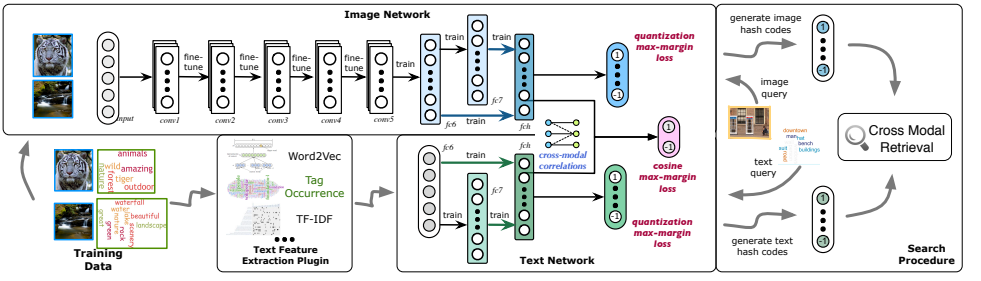


Figure 1: Correlation Hashing Network (CHN) for cross-modal retrieval, which constitutes (1) a convolutional network (CNN) for learning image representations, (2) a multilayer perceptrons (MLP) for learning text representations, (3) two hashing layers fch for generating hash codes, (4) a cosine max-margin loss for capturing cross-modal correlations, and a quantization max-margin loss for controlling hashing quality.

between the fch-layer representations $\boldsymbol{u}_i, \boldsymbol{v}_i$ and the binary hash codes $\boldsymbol{h}_i^x, \boldsymbol{h}_i^y$. We design new loss functions over hash codes generated by the deep networks for cross-modal correlation learning and quantization error minimization, which enable effective cross-modal retrieval.

3.2 Cosine Max-Margin Loss

For a pair of binary codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, there is a relationship between their Hamming distance $\text{dist}_H(\cdot,\cdot)$ and their inner product $\langle\cdot,\cdot\rangle$: $\text{dist}_H\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \frac{1}{2}\left(b - \left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right)$. Thus, we may use the inner product as a reasonable surrogate of the Hamming distance to quantify the pairwise similarity. However, note that $\boldsymbol{h}_i^x = \text{sgn}(\boldsymbol{u}_i)$ and $\boldsymbol{h}_i^y = \text{sgn}(\boldsymbol{v}_i)$, hence the approximation of such a surrogate for continuous representations $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ will be inaccurate if their vector lengths are very different, i.e. $\frac{1}{2}\left(b - \langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle\right) \in (-\infty, +\infty)$ will no longer be a good surrogate of $\text{dist}_H\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) \in [-b, +b]$. Figure 2 shows such a bad case, where points 1 and 2 (in red) have very different vector lengths and thus large Euclidean distance, but their Hamming distance is 0 since they are assigned to the same binary code $(1,-1,1)$. The gap between Hamming distance and inner product has raised a serious misspecification issue of existing inner product based deep hashing methods [15, 32].

To close the gap between Hamming distance and inner product for continuous representations, note that for a pair of binary codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, there is another relationship between their Hamming distance $\text{dist}_H(\cdot,\cdot)$ and the cosine distance $\cos(\cdot,\cdot)$: $\text{dist}_H\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \frac{b}{2}\left(1 - \cos\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)\right)$, where $\cos(\boldsymbol{u}_i, \boldsymbol{v}_j) = \frac{\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle}{\|\boldsymbol{u}_i\| \|\boldsymbol{v}_j\|}$, and $\|\cdot\|$ is the vector length. Since cosine distance can mitigate the diversity of vector lengths and make the continuous representations $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ lie on the unit sphere (which is important for cross-modal data as they usually have very different vector lengths), it makes $\frac{b}{2}(1 - \cos(\boldsymbol{u}_i, \boldsymbol{v}_j)) \in [-b, +b]$ a more accurate surrogate of $\text{dist}_H\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right)$ especially for comparing continuous representations of different modalities. As can be seen in Figure 2, the cosine distance between points 1 and 2 (in red) is close to zero and thus better approximates their Hamming distance. Hence in this paper, we opt to use the cosine distance as a good surrogate of the Hamming distance, which

leads to new cosine-distance based structural loss functions.

To maximize the cross-modal correlation, we propose the following criterion: for each pair of objects $(\boldsymbol{x}_i, \boldsymbol{y}_j, s_{ij})$, if $s_{ij} = 1$, indicating that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are similar, then their binary hash codes must be similar across different modalities, i.e. the Hamming distance should satisfy $d_H(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y) \to 0$, which implies the cosine distance should satisfy $\cos(\boldsymbol{u}_i, \boldsymbol{v}_j) \to 1$. Correspondingly, if $s_{ij} = -1$, indicating that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are dissimilar, then by derivation, the cosine distance should satisfy $\cos(\boldsymbol{u}_i, \boldsymbol{v}_j) \to -1$. It is very important to note that, for other widely-used distance metrics (e.g. inner product, Euclidean distance, etc), it is very difficult to devise such a well-specified learning criterion because these distances are not good surrogates of the Hamming distance. A straight-forward loss for achieving the above goal is the squared loss $(s_{ij} - \cos(\boldsymbol{u}_i, \boldsymbol{v}_j))^2$, however, the squared loss is not robust to outlier pairs of points. Motivated by SVMs, the similarity-preserving criterion leads to a novel cosine max-margin loss for maximizing cross-modal correlation as

$$L = \sum_{s_{ij} \in \mathcal{S}} \max\left(0, \delta - s_{ij} \frac{\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle}{\|\boldsymbol{u}_i\| \, \|\boldsymbol{v}_j\|}\right)^2, \tag{1}$$

where $0 < \delta \leq 1$ is the margin parameter. The range of cosine distance $\cos(\boldsymbol{u}_i, \boldsymbol{v}_j) \in [-1, 1]$ is consistent with binary similarity labels $s_{ij} \in \{-1, 1\}$, making the cosine max-margin loss in Equation (1) a well-specified loss for preserving the pairwise similarity information conveyed in $\mathcal{S}$. The cosine max-margin loss loss is powerful for cross-modal correlation analysis, since the vector lengths are very diverse in different modalities and may make other distance metrics (e.g. inner product) misspecified. In real retrieval systems, cosine distance is widely used to mitigate the diversity of vector lengths and significantly improve retrieval quality, but to date, it has not been explored in deep hashing for cross-modal retrieval [28].

3.3 Quantization Max-Margin Loss

Though we justify that cosine distance is a good surrogate of Hamming distance, such an approximation may fail when two similar points $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ with $s_{ij} = 1$ (i.e. their cosine distance is small due to minimizing the cosine max-margin loss) lie on different sides of the hyperplane (i.e. their Hamming distance is large due to different signs of hash codes across the hyperplane). Figure 2 shows such a failure case, where points 3 and 4 (in purple) have small cosine distance but large Hamming distance because they are assigned with different binary codes $(1, -1, 1)$ and $(1, 1, 1)$, respectively. This contradiction makes cosine distance an inaccurate surrogate of Hamming distance when the points are near the splitting-plane (e.g. $y = 0$, in purple). Such a gap between Hamming distance and cosine distance should be reduced for better Hamming approximation.

To close the gap between Hamming distance and cosine distance under continuous representations, note that for a pair of continuous representations $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, if they are close (in cosine distance) to their signed codes $\boldsymbol{h}_i^x = \operatorname{sgn}(\boldsymbol{u}_i)$ and $\boldsymbol{h}_i^y = \operatorname{sgn}(\boldsymbol{v}_i)$ (i.e. far from the hyperplane), then they will lie in the same hypercube with high probability (i.e. with the same binary code and hence their Hamming distance is zero). As shown in Figure 2, we want points 3 and 4 to lie near the vertex of the hypercube (e.g. $(1, 1, 1)$) and far from the splitting-plane $y = 0$ (in purple). That is, we should favor points 5 and 6 and prevent points 3 and 4. To this end, instead of using the squared loss $(1 - \cos(|\boldsymbol{u}_i|, \mathbf{1}))^2$ which is not robust to outlier bits especially for unbalanced encoding, we minimizes the inconsistency between

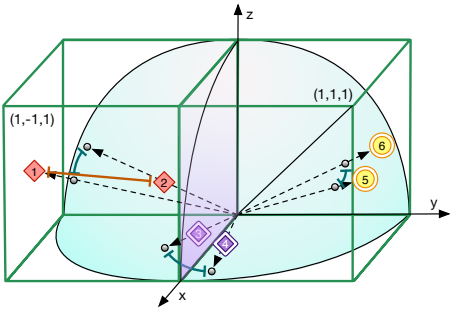


Figure 2: Motivation of the cosine max-margin loss and the quantization max-margin loss. (1) Similar points 1 and 2 (in red): large Euclidean distance (bad Hamming surrogate) but small cosine distance (better Hamming surrogate). (2) Similar points 3 and 4 (in purple): small cosine distance but large Hamming distance (the gap between cosine and Hamming). (3) Similar points 5 and 6 (in yellow): small cosine distance and small Hamming distance (the gap between cosine and Hamming is closed by the quantization max-margin loss).

cosine distance and Hamming distance by proposing a new quantization max-margin loss

$$Q = \sum_{i=1}^{n_x} \max\left(0, \delta - \frac{\langle |\boldsymbol{u}_i|, \mathbf{1}\rangle}{\|\boldsymbol{u}_i\| \, \|\mathbf{1}\|}\right) + \sum_{i=1}^{n_y} \max\left(0, \delta - \frac{\langle |\boldsymbol{v}_i|, \mathbf{1}\rangle}{\|\boldsymbol{v}_i\| \, \|\mathbf{1}\|}\right), \tag{2}$$

where $0 < \delta \leq 1$ is the margin parameter. Note that, minimizing the quantization max-margin loss jointly with minimizing the cosine max-margin loss will not only close the gap between the Hamming distance and cosine distance, but also lead to lower quantization error when binarizing the continuous representations $\boldsymbol{u}_i \in \mathbb{R}^b$ and $\boldsymbol{v}_j \in \mathbb{R}^b$ to hash codes $\boldsymbol{h}_i^x = \text{sgn}(\boldsymbol{u}_i) \in \{-1,1\}^b$ and $\boldsymbol{h}_j^y = \text{sgn}(\boldsymbol{v}_j) \in \{1,-1\}^b$, especially for unbalanced codes with outlier bits.

3.4 Hash Function Learning

We perform end-to-end representation learning and hash encoding by integrating Equations (1)–(2) in a joint optimization problem

$$\min_{\Theta} O \triangleq L + \lambda Q, \tag{3}$$

where $\Theta \triangleq \{\boldsymbol{W}^\ell, \boldsymbol{b}^\ell\}$ is the set of network parameters, λ is the tradeoff parameter for the quantization max-margin loss. Through problem (3), we can achieve optimal hash codes for efficient cross-modal retrieval. Finally, we obtain b-bit binary codes by binarization as $\boldsymbol{h}_x \leftarrow \text{sgn}(\boldsymbol{u})$ and $\boldsymbol{h}_y \leftarrow \text{sgn}(\boldsymbol{v})$, where $\forall i$, $\text{sgn}(u_i) = 1$ if $u_i > 0$, otherwise $\text{sgn}(u_i) = -1$. Since we have minimized the quantization max-margin error in (3) during training, this final binarization step will incur very small sacrifice of retrieval quality as validated empirically.

3.5 Learning Algorithm

We derive learning algorithms for CHN in Equation (3), and show rigorously that both cosine max-margin loss and quantization max-margin loss can be optimized efficiently via

standard back-propagation (BP) algorithm. For brevity, we define the pointwise cost of the image modality (the pointwise cost of the text modality is the same and omitted) as $O_i^x \triangleq \sum_{j:s_{ij}\in\mathcal{S}} L_{ij} + \lambda Q_i^x$. We derive the gradient of point-wise cost O_i^x w.r.t. $\boldsymbol{W}_{x,k}^{\ell}$, the network parameter of the k-th unit in the ℓ-th layer for the image network as

$$\frac{\partial O_i^x}{\partial \boldsymbol{W}_{x,k}^{\ell}} = \sum_{j:s_{ij}\in\mathcal{S}} \frac{\partial L_{ij}}{\partial \boldsymbol{W}_{x,k}^{\ell}} + \lambda \frac{\partial Q_i^x}{\partial \boldsymbol{W}_{x,k}^{\ell}} = \left(\sum_{j:s_{ij}\in\mathcal{S}} \frac{\partial L_{ij}}{\partial \hat{u}_{ik}^{\ell}} + \lambda \frac{\partial Q_i^x}{\partial \hat{u}_{ik}^{\ell}} \right) \frac{\partial \hat{u}_{ik}^{\ell}}{\partial \boldsymbol{W}_{x,k}^{\ell}} = \delta_{x,ik}^{\ell} \boldsymbol{u}_i^{\ell-1}, \tag{4}$$

where $\hat{\boldsymbol{u}}_i^{\ell} = \boldsymbol{W}_x^{\ell} \boldsymbol{u}_i^{\ell-1} + \boldsymbol{b}_x^{\ell}$ is ℓ-th layer output before activation $a_x^{\ell}(\cdot)$, $\delta_{x,ik}^{\ell} \triangleq \sum_{j:s_{ij}\in\mathcal{S}} \frac{\partial L_{ij}}{\partial \hat{u}_{ik}^{\ell}} + \lambda \frac{\partial Q_i^x}{\partial \hat{u}_{ik}^{\ell}}$ is the point-wise *residual* term that measures how much the k-th unit in the ℓ-th layer is responsible for the error of point $\boldsymbol{x}_i$ in the network output. For an output unit k, we can measure the difference between the network's activation and the true target value, and use that to define the residual $\delta_{x,ik}^{l}$ as

$$\begin{aligned} \delta_{x,ik}^{l} = & \sum_{j:s_{ij}\in\mathcal{S}} 2 \cdot \max\left(0, \delta - s_{ij} \frac{\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle}{\|\boldsymbol{u}_i\| \|\boldsymbol{v}_j\|}\right) \cdot \mathbb{I}\left(\delta - s_{ij} \frac{\boldsymbol{u}_i^l \cdot \boldsymbol{v}_j^l}{\|\boldsymbol{u}_i^l\| \|\boldsymbol{v}_j^l\|} > 0\right) \dot{a}_x^l(\hat{u}_{ik}^l) \cdot \left[-s_{ij} \left(\frac{v_{jk}^l}{\|\boldsymbol{u}_i^l\| \|\boldsymbol{v}_j^l\|} - \frac{u_{ik}^l \left\langle \boldsymbol{u}_i^l, \boldsymbol{v}_j^l \right\rangle}{\|\boldsymbol{u}_i^l\|^3 \|\boldsymbol{v}_j^l\|} \right) \right] \\ & - \lambda \dot{a}_x^l(\hat{u}_{ik}^l) \cdot \mathbb{I}\left(\delta - \frac{\sum_{j=1}^b \left| u_{ij}^l \right|}{\sqrt{b} \|\boldsymbol{u}_i^l\|} > 0\right) \left[\frac{\operatorname{sgn}(u_{ik}^l)}{\sqrt{b} \|\boldsymbol{u}_i^l\|} - \frac{u_{ik}^l \sum_{j=1}^b \left| u_{ij}^l \right|}{\sqrt{b} \|\boldsymbol{u}_i^l\|^3} \right] \end{aligned} \tag{5}$$

where $\dot{a}_x^l(\cdot)$ is the derivative of the l-th layer activation function, and $\mathbb{I}(A)$ is an indicator function, $\mathbb{I}(A) = 1$ if A is true and $\mathbb{I}(A) = 0$ otherwise. For a hidden unit k in the $(\ell-1)$-th layer, we compute residual $\delta_{x,ik}^{\ell-1}$ based on a weighted average of the errors of all units $k' = 1, \ldots, n_{\ell-1}$ in the $(\ell-1)$-th layer that use $\boldsymbol{u}_i^{\ell-1}$ as an input, which is consistent with BP,

$$\delta_{x,ik}^{\ell-1} = \left(\sum_{k'=1}^{n_{\ell-1}} \delta_{x,ik'}^{\ell} W_{x,kk'}^{\ell-1} \right) \dot{a}_x^{\ell-1}\left(\hat{u}_{ik}^{\ell-1}\right), \tag{6}$$

where $n_{\ell-1}$ is number of units in $(\ell-1)$-th layer. The residuals in all other layers can be computed by back-propagation. The overall computational complexity is $O(|\mathcal{S}|)$, where $|\mathcal{S}|$ is the number of cross-modal similarity pairs in $\mathcal{S}$ for training.

A nice property of the proposed algorithm is that it only requires computing the residual of the output layer involves the pairwise summation as in Equation (5). For all hidden layers, all the residuals can be simply computed recursively by Equation (6), which does not involve pairwise summation. Hence we do not need to modify standard BP in all hidden layers $1 \le \ell \le l-1$ but only modify standard BP by replacing the output residual with Equation (5).

4 Experiments

4.1 Setup

NUS-WIDE [8] is a web image dataset of 81 ground truth concepts manually annotated for evaluation. Following prior works [4, 29], we use the subset of 195,834 image-text pairs that belong to some of the 21 most frequent concepts. All images are resized into 256×256.
MIR-Flickr [10] consists of 25,000 images collected from the Flickr website, where each image is labeled with some of 38 semantic concepts. All images are resized into 256×256.

For our deep learning based approach CHN, we directly use the raw image pixels as the input. For fair comparison, for traditional shallow hashing methods, we use AlexNet [13] to

Table 1: Mean Average Precision (MAP) on Single-Modal Retrieval Task ($I \rightarrow I$)

Task	Method	NUS-WIDE				MIR-Flickr			
		12 bits	24 bits	32 bits	48 bits	12 bits	24 bits	32 bits	48 bits
$I \rightarrow I$	DNNH	0.674	0.697	0.713	0.715	0.783	0.789	0.791	0.802
	DHN	0.708	0.735	0.748	0.758	0.810	0.828	0.829	0.841
	CHN	**0.718**	**0.745**	**0.760**	**0.768**	**0.817**	**0.829**	**0.843**	**0.849**

extract deep *fc7* features for each image in two benchmark datasets by a 4096-dimensional vector. For text modality, all the methods use tag occurrence vectors as the input. 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. In MIR-Flickr, we randomly select 1000 pairs as the query set, 4000 pairs as the training set and 1000 pairs as the validation set. The similarity pairs for training are constructed using semantic labels: each pair is similar (dissimilar) if they share at least one (none) semantic label. We compare CHN with state-of-the-art cross-modal hashing and deep hashing methods, including three unsupervised methods **IMH** [25], **CVH** [14], and **MMNN** [21], and five supervised methods **CMSSH** [1], **RaHH** [22], **SCM** [34], **SePH** [17] and **DCMH** [12], where **MMNN** and **DCMH** are deep hashing methods. We follow [12, 34] to evaluate retrieval quality via three standard metrics: Mean Average Precision (MAP), precision-recall curves and precision@top-R curves.

We implement the CHN model based on the open-source **Caffe** framework [11]. For image network, we employ the AlexNet [13], fine-tune $conv1-fc7$ that were copied from the pre-trained model, and train hashing layer fch, all via back-propagation. For text network, we employ a two layer multilayer perceptrons (MLP), in which the $fc7$ layer has 4096 ReLU units with dropout rate 0.5, and the fch layer have b tanh units. We use mini-batch SGD with 0.9 momentum, and cross-validate the learning rate from 10^{-5} to 1 with a multiplicative step-size 10, and fix mini-batch size as 64. For all methods, we select parameters via cross-validation. Each experiment repeats five runs and average results are reported.

4.2 Results

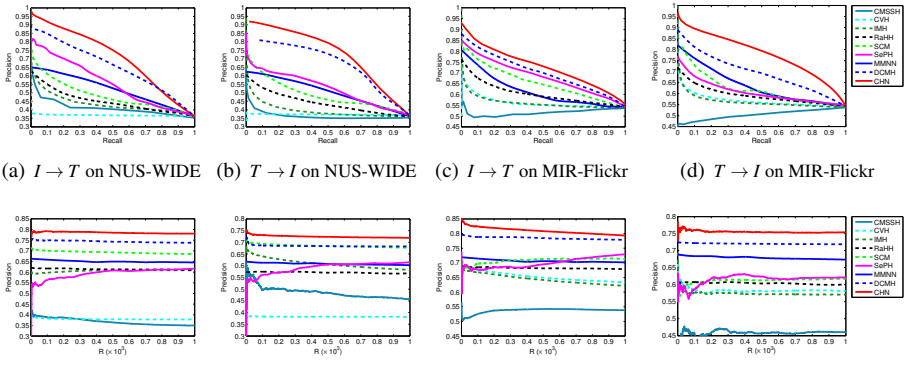
We report in Table 2 the MAP of all methods with different code lengths, i.e. 16, 32, 64 and 128 bits. CHN substantially outperforms all state-of-the-art methods for all cross-modal retrieval tasks. Specifically, for NUS-WIDE dataset, CHN outperforms the best shallow method SCM by 9.19% / 6.44% in average MAP for $I \rightarrow T$ / $T \rightarrow I$. For MIR-Flickr dataset, CHN outperforms the best shallow method SePH by 9.74% / 15.25% in average MAP for $I \rightarrow T$ / $T \rightarrow I$. Compared to deep cross-modal hashing methods, CHN outperforms state-of-the-art DCMH by large margins of 6.15% / 6.49% and 5.51% / 6.53%. These results verify that CHN is able to learn high-quality hash codes for effective cross-modal retrieval.

We respectively report in Figure 3 (a)-(d) the precision-recall curves with 32 bits for two cross-modal retrieval tasks $I \rightarrow T$ and $T \rightarrow I$ on two benchmark datasets NUS-WIDE and MIR-Flickr. CHN shows the best retrieval performance at all recall levels. Figure 3 (e)-(h) respectively show the precision@top-R curves of all state-of-the-art methods, which further represent the precision changes along with the number of top-R retrieved results ($R = 1000$) with 32 bits on NUS-WIDE and MIR-Flickr datasets. CHN significantly outperforms all state-of-the-art methods under these metrics.

To verify the effectiveness of our proposed CHN approach, we slightly adapt CHN to support single-modal retrieval task ($I \rightarrow I$) via rewriting (1) with $L = \sum_{s_{ij} \in \mathcal{S}} \max\left(0, \delta - s_{ij} \frac{\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle}{\|\boldsymbol{u}_i\| \|\boldsymbol{u}_j\|}\right)^2$

Table 2: Comparison of MAP on Two Cross-Modal Retrieval Tasks ($I \rightarrow T$ and $T \rightarrow I$)

Task	Method	NUS-WIDE				MIR-Flickr			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
$I \rightarrow T$	CVH [24]	0.4454	0.4342	0.4290	0.4479	0.6883	0.7092	0.6976	0.6334
	IMH [25]	0.5256	0.6358	0.6151	0.6183	0.6765	0.6989	0.6964	0.6839
	CMSSH [1]	0.4665	0.4809	0.5670	0.5288	0.5122	0.5404	0.5842	0.5740
	RaHH [27]	0.6047	0.6312	0.6354	0.6534	0.6899	0.7086	0.7155	0.7204
	SCM [34]	0.6871	0.7271	0.7600	0.7739	0.6953	0.7091	0.7070	0.7497
	SePH [17]	0.5982	0.5910	0.5988	0.6239	0.7526	0.7604	0.7607	0.7651
	MMNN [21]	0.6255	0.6424	0.6514	0.6713	0.6915	0.7185	0.7277	0.7352
	DCMH [12]	0.7353	0.7628	0.7805	0.7912	0.7576	0.7985	0.8152	0.8369
	CHN	**0.7995**	**0.8146**	**0.8353**	**0.8662**	**0.8223**	**0.8477**	**0.8777**	**0.8808**
$T \rightarrow I$	CVH [24]	0.4357	0.4253	0.4186	0.4184	0.6065	0.6277	0.6063	0.6004
	IMH [25]	0.6253	0.6816	0.7094	0.6532	0.6229	0.6201	0.6239	0.6237
	CMSSH [1]	0.4166	0.5110	0.4343	0.4974	0.4656	0.4624	0.4769	0.5337
	RaHH [27]	0.5786	0.6158	0.6214	0.6240	0.6248	0.6321	0.6359	0.6464
	SCM [34]	0.6794	0.7194	0.7480	0.7466	0.6173	0.6115	0.6177	0.6564
	SePH [17]	0.6044	0.6036	0.6256	0.6405	0.6470	0.6429	0.6517	0.6550
	MMNN [21]	0.6083	0.6226	0.6435	0.6648	0.6815	0.6992	0.7082	0.7171
	DCMH [12]	0.6898	0.7102	0.7358	0.7557	0.7013	0.7288	0.7458	0.7698
	CHN	**0.7533**	**0.7803**	**0.7888**	**0.8288**	**0.7749**	**0.7891**	**0.8169**	**0.8258**



(a) $I \rightarrow T$ on NUS-WIDE (b) $T \rightarrow I$ on NUS-WIDE (c) $I \rightarrow T$ on MIR-Flickr (d) $T \rightarrow I$ on MIR-Flickr

(e) $I \rightarrow T$ on NUS-WIDE (f) $T \rightarrow I$ on NUS-WIDE (g) $I \rightarrow T$ on MIR-Flickr (h) $T \rightarrow I$ on MIR-Flickr

Figure 3: Precision-recall curves (a)-(d) and Precision@top R curves (e)-(h) @ 32 bits.

and (2) with $Q = \sum_{i=1}^{n} \max\left(0, \delta - \frac{\langle \boldsymbol{u}_i, \mathbf{1} \rangle}{\|\boldsymbol{u}_i\| \|\mathbf{1}\|}\right)$. We compare with the state-of-the-art single-modal deep hashing methods, DHN [37] and DNNH [15], following the evaluation protocols in [37] [15] on the two benchmark datasets NUS-WIDE and MIR-Flickr. The MAP results w.r.t different lengths of bits, i.e. 12, 24, 32 and 48 bits, are shown in Table 1. From this table, we can observe that the proposed CHN outperforms the state-of-the-art single-modal deep hashing method DHN by 1.05% / 0.75% on NUS-WIDE and MIR-Flickr, respectively. This validates that CHN can generate high-quality hash codes for effective single-modal retrieval.

4.3 Discussion

To go deeper with the efficacy of CHN, we investigate four variants of CHN: **CHN-M** is the CHN variant without using the margin parameter, in other words, $\delta = 1.0$; **CHN-I** is the CHN variant that replaces the cosine max-margin loss (1) with the widely-used inner-product

Table 3: Mean Average Precision (MAP) of CHN Variants on NUS-WIDE and MIR-Flickr

Task	Method	NUS-WIDE				MIR-Flickr			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
$I \to T$	CHN-M	0.6987	0.7251	0.7350	0.7593	0.7313	0.7665	0.8148	0.8281
	CHN-I	0.6751	0.7047	0.7214	0.7253	0.7105	0.7481	0.7798	0.7843
	CHN-Q	0.7743	0.7982	0.8212	0.8598	0.7893	0.8214	0.8558	0.8678
	CHN	0.7995	0.8146	0.8353	0.8662	0.8223	0.8477	0.8777	0.8808
	CHN-B	**0.8650**	**0.8706**	**0.8746**	**0.8879**	**0.8753**	**0.8612**	**0.8905**	**0.8933**
$T \to I$	CHN-M	0.5789	0.5924	0.5997	0.6318	0.6532	0.6875	0.7015	0.7135
	CHN-I	0.5874	0.6012	0.6241	0.6534	0.6817	0.6987	0.7314	0.7389
	CHN-Q	0.7395	0.7543	0.7779	0.8053	0.7515	0.7687	0.8043	0.8125
	CHN	0.7533	0.7803	0.7888	0.8288	0.7749	0.7891	0.8169	0.8258
	CHN-B	**0.8003**	**0.8095**	**0.8185**	**0.8407**	**0.7915**	**0.8142**	**0.8207**	**0.8308**

squared loss $L = \sum_{s_{ij} \in \mathcal{S}} \left(s_{ij} - \frac{1}{b} \left\langle \boldsymbol{h}_i, \boldsymbol{h}_j \right\rangle\right)^2$ [18, 32]; **CHN-Q** is the CHN variant without using the quantization max-margin loss (2); **CHN-B** is the CHN variant without using binarization on hash codes, which may serve as the upper bound of retrieval performance.

From Table 3, we have the following key observations. **(a)** CHN outperforms CHN-M by large margins, demonstrating that the max-margin principle can significantly enhance the robustness of the hash codes to the outlier points. **(b)** By using the cosine max-margin loss, CHN outperforms CHN-I by large margins. The squared inner-product loss has been widely adopted in the previous works [18, 32]. However, this loss cannot link well the pairwise distances between continuous representations (taking values in $(-\infty, +\infty)$ when using continuous relaxation) to the pairwise similarity labels (taking binary values {-1,1}). In contrast, the proposed cosine max-margin loss (1) is inherently consistent with the training pairs. Besides, the margin parameter δ can also control the robustness level of the similarity-preserving procedure to the outlier points. The promising performance of CHN suggests that the proposed cosine max-margin loss can preserve cross-modal correlations and is well-specified to cross-modal retrieval scenarios. **(c)** By using quantization max-margin loss (2), CHN incurs small MAP decreases than CHN-Q when quantizing continuous representations into binary codes. Especially for shorter length of hash codes (16 bits), CHN-Q incurs huge decreases while CHN incurs negligible MAP decreases. This validates that quantization max-margin loss can effectively reduce the quantization error and obtain high-quality hash codes. The results also imply that all components in CHN are vital for achieving the promising performance, and missing any component will lead to huge performance drop of cross-modal retrieval.

5 Conclusion

In this paper, we have proposed a novel Correlation Hashing Network (CHN) for effective and efficient cross-modal retrieval. CHN is a hybrid deep architecture that jointly optimizes the new cosine max-margin loss on semantic similarity pairs and the new quantization max-margin loss on compact hash codes. Experiments on standard cross-modal retrieval datasets show that CHN model yields substantial boosts over state-of-the-art hashing methods.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2016YFB1000701), the National Natural Science Foundation of China (No. 61502265, 61325008, and 71690231), the National S&T Supporting Program (2015BAF32B01), and Tsinghua TNList Key Projects.

References

[1] M.M. Bronstein, A.M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*. IEEE, 2010.

[2] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454. ACM, 2016.

[3] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, and Qingfu Wen. Deep quantization network for efficient image retrieval. In *AAAI*, 2016.

[4] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, pages 3974–3980, 2017.

[5] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Deep visual-semantic quantization for efficient image retrieval. In *CVPR*, 2017.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*. ACM, 2009.

[7] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[8] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*. IEEE, 2011.

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

[10] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *ICMR*. ACM, 2008.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia Conference*. ACM, 2014.

[12] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. *CoRR*, abs/1602.02255, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

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

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

[16] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations (ICLR), 2014 (arXiv:1409.1556)*, 2014.

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

[18] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *CVPR*. IEEE, 2012.

[19] Xianglong Liu, Junfeng He, Cheng Deng, and Bo Lang. Collaborative hashing. In *CVPR*. IEEE, 2014.

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu. Composite correlation quantization for efficient multimodal retrieval. *SIGIR*, 2016.

[21] Jonathan Masci, Michael M. Bronstein, Alexander M. Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *IEEE TPAMI*, 36, 2014.

[22] Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, Wenwu Zhu, and Shiqiang Yang. Comparing apples to oranges: a scalable solution with heterogeneous hashing. In *SIGKDD*. ACM, 2013.

[23] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36, 2014.

[24] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22, 2000.

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

[26] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 2014.

[27] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *MM*. ACM, 2014.

[28] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. Arxiv, 2014.

[29] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. In *VLDB*. ACM, 2014.

[30] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *NIPS*, 2009.

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

[32] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *AAAI*, 2014.

[33] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *SIGIR*. ACM, 2014.

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

[35] Yi Zhen and Dit-Yan Yeung. A probabilistic model for multimodal hash function learning. In *SIGKDD*. ACM, 2012.

[36] Yi Zhen and Dit-Yan Yeung. Co-regularized hashing for multimodal data. In *NIPS*, 2012.

[37] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao. Deep hashing network for efficient similarity retrieval. In *AAAI*, 2016.