Collective Deep Quantization for Efficient Cross-Modal Retrieval*

Yue Cao, Mingsheng Long, Jianmin Wang, Shichen Liu
KLiss, MOE; TNList; School of Software, Tsinghua University, Beijing, China
{caoyue10,liushichen95}@gmail.com {mingsheng,jimwang}@tsinghua.edu.cn

Abstract

Cross-modal similarity retrieval is a problem about designing a retrieval system that supports querying across content modalities, e.g., using an image to retrieve for texts. This paper presents a compact coding solution for efficient cross-modal retrieval, with a focus on the quantization approach which has already shown the superior performance over the hashing solutions in single-modal similarity retrieval. We propose a collective deep quantization (CDQ) approach, which is the first attempt to introduce quantization in end-to-end deep architecture for cross-modal retrieval. The major contribution lies in jointly learning deep representations and the quantizers for both modalities using carefully-crafted hybrid networks and well-specified loss functions. In addition, our approach simultaneously learns the common quantizer codebook for both modalities through which the cross-modal correlation can be substantially enhanced. CDQ enables efficient and effective cross-modal retrieval using inner product distance computed based on the common codebook with fast distance table lookup. Extensive experiments show that CDQ yields state of the art cross-modal retrieval results on standard benchmarks.

Introduction

While multimedia big data with large volumes and high dimensions are pervasive in search engines and social networks, it has attracted increasing attention to enable approximate nearest neighbors (ANN) search across different media modalities with both 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 and quantization (Wang et al. 2014a), which transform high-dimensional data into compact binary codes and generate similar binary codes for similar data items. Due to large volumes and high dimensions, effective and efficient cross-modal retrieval remains a challenge.

In this paper, we are interested in the cross-modal quantization approach that represents each point by a short code formed by the index of the nearest center, as quantization (Ge et al. 2014; Zhang, Du, and Wang 2014) has shown more powerful representation ability than hashing in single-modal search. It has been widely studied in cross-modal similarity search with typical solutions based on hashing (Wang et al. 2014a; Bronstein et al. 2010; Kumar and Udupa 2011; Song et al. 2013), while relatively unexplored in cross-modal search except only two quantization approaches (Long et al. 2016; Zhang and Wang 2016). However, without learning deep representations, existing cross-modal quantization approaches cannot close the gap across different modalities.

Recent deep hashing methods (Xia et al. 2014; Lai et al. 2015) show that both feature representation and hash coding can be learned more effectively using deep networks (Krizhevsky, Sutskever, and Hinton 2012; Lin, Chen, and Yan 2014), which can naturally encode nonlinear hashing functions. Other cross-modal retrieval models via deep learning (Masci et al. 2014; Jiang and Li 2016; Cao, Long, and Wang 2016) have shown that deep models can capture nonlinear cross-modal correlations more effectively and yielded state-of-the-art results on many benchmarks. However, without exploring the quantization techniques, existing deep hashing methods cannot minimize the quantization error to generate high-quality binary codes. Furthermore, deep features may not be quantized effectively using post-step quantization techniques—if deep features do not exhibit a cluster structure, then they may not be quantized accurately (Ge et al. 2014). Hence it is important to improve the quantizability of the deep representations in an end-to-end architecture such that they can be quantized more effectively.

This paper presents Collective Deep Quantization (CDQ), which is the first attempt to cross-modal deep quantization. CDQ jointly learns deep image and text representations tailored to binary 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 bottleneck layers for learning quantizable representations, (3) an adaptive cross-entropy loss for capturing cross-modal correlations, and (4) a collective quantization loss for controlling coding quality and the quantizability of representations. Extensive experiments show that CDQ yields state of the art cross-modal retrieval performance.

*Corresponding authors: Mingsheng Long and Jianmin Wang.

Related Work

Cross-modal hashing has been a popular research topic in machine learning, computer vision, and multimedia retrieval (Bronstein et al. 2010; Kumar and Udupa 2011; Zhen and Yeung 2012a; 2012b; Wang et al. 2014b; Yu et al. 2014; Hu et al. 2014; Zhang and Li 2014; Liu et al. 2014; Cao et al. 2016c; Long et al. 2016; Cao et al. 2016a). We refer readers to (Wang et al. 2014a) for a comprehensive survey.

Prior cross-modal hashing methods can be categorized into unsupervised methods and supervised methods. IMH (Song et al. 2013) and CVH (Kumar and Udupa 2011) are unsupervised methods that extend spectral hashing (Weiss, Torralba, and Fergus 2009) to multimodal data. CMSSH (Bronstein et al. 2010), SCM (Zhang and Li 2014) and SePH (Lin et al. 2015) 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. Prior cross-modal hashing methods based on shallow architectures cannot effectively exploit the correlations across different modalities. Deep multimodal embedding methods (Frome et al. 2013; Kiros, Salakhutdinov, and Zemel 2014; Donahue et al. 2015; Gao et al. 2015) have shown that deep models can bridge heterogeneous modalities more effectively. Recent deep hashing methods (Lai et al. 2015; Zhu et al. 2016; Cao et al. 2016b) have given state of the art results on many image datasets, but they can only be used for single-modal retrieval.

There is a few previous work closely related to our work. Deep Cross-Modal Hashing (DCMH) (Jiang and Li 2016) and Correlation Hashing Network (CHN) (Cao, Long, and Wang 2016) are the only two cross-modal deep hashing methods that use deep convolutional networks (Krizhevsky, Sutskever, and Hinton 2012) for image representation and multilayer perceptrons (Rumelhart, Hinton, and Williams 1986) for text representation. However, neither DCMH nor CHN explores the quantization technique to minimize the quantization error and improve the quantizability of deep representations. Hence, they may produce suboptimal binary codes for cross-modal retrieval. Composite Correlation Quantization (CCQ) (Long et al. 2016) and Collaborative Quantization (Zhang and Wang 2016) are the only two cross-modal quantization methods which achieve much better performance than cross-modal hashing methods. However, they cannot learn deep representations to close the gap across different modalities. This work addresses these issues by proposing a novel cross-modal deep quantization method.

Collective Deep Quantization

In cross-modal retrieval, the database consists of objects from one modality and the query consists of objects from another modality. We maximize the correlation underlying different modalities by learning from a training set of N_x images $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and N_y texts $\{\mathbf{y}_j\}_{j=1}^{N_y}$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ denotes the D_x-dimensional feature vector of the image modality, and $\mathbf{y}_j \in \mathbb{R}^{D_y}$ 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 $\mathbf{x}_i$ and $\mathbf{y}_j$ are similar and $s_{ij} = 0$ indicates $\mathbf{x}_i$ and $\mathbf{y}_j$ 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 CDQ is to jointly learn modality-specific quantizers $f_x(\mathbf{x}) : \mathbb{R}^{D_x} \mapsto \{0,1\}^B$ and $f_y(\mathbf{y}) : \mathbb{R}^{D_y} \mapsto \{0,1\}^B$ which encode each unimodal point $\mathbf{x}$ and $\mathbf{y}$ in compact B-bit binary code $\mathbf{b}_x = f_x(\mathbf{x})$ and $\mathbf{b}_y = f_y(\mathbf{y})$ such that the similarity information conveyed in the given bimodal object pairs $\mathcal{S}$ is maximally preserved.

Collective Deep Quantization (CDQ) is a hybrid deep architecture for supervised learning to hash, as shown in Figure 1. The hybrid architecture accepts input in a pairwise form $(\mathbf{x}_i, \mathbf{y}_j, s_{ij})$ and processes them through deep representation learning and binary coding pipeline: (1) an image network with multiple convolution-pooling layers to extract good image representations, and a text network with several fully-connected layers to extract good text representations; (2) two fully-connected bottleneck layers to generate optimal dimension-reduced representations; (3) an adaptive cross-entropy loss for capturing cross-modal correlations; and (4) a collective quantization loss for controlling hashing quality and the quantizability of bottleneck representations.

Model Formulation

The hybrid deep architecture for learning cross-modal quantizers are shown in Figure 1, which constitutes an image network and a text network. In the image network, we extend AlexNet (Krizhevsky, Sutskever, and Hinton 2012), 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 fcb bottleneck layer with R hidden units, which transforms the $fc7$ layer representation to R-dimensional bottleneck representation $\mathbf{z}_i^x$. In text network, we adopt the Multilayer perceptrons (MLP) (Rumelhart, Hinton, and Williams 1986) comprising three fully connected layers, of which the last layer is replaced with a new fcb bottleneck layer with R hidden units, which transforms the network activation to R-dimensional bottleneck representation $\mathbf{z}_j^y$. To encourage the fcb layer representation $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ to be quantizable for binary coding, we use the hyperbolic tangent (tanh) activation function $a(\mathbf{z}) = \tanh(\mathbf{z})$ to produce nonlinear dimension-reduced representation. Several well-specified loss functions are added on top of the hybrid deep network for cross-modal correlation learning and quantization error minimization to enable effective and efficient cross-modal retrieval.

In this paper, we guarantee that the fcb representation $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ will be optimal for compact binary coding by jointly (1) preserving the similarity between given pairs in $\mathcal{S}$, (2) controlling the quantization error of binarizing the fcb representation $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ into binary codes $\mathbf{b}_i^x$ and $\mathbf{b}_j^y$, and (3) improving the quantizability of the fcb representation $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ such that it can be quantized effectively. These goals can be implemented in a Bayesian learning framework as follows. For a pair of objects $\mathbf{x}_i$ and $\mathbf{y}_j$, we can use the inner product on their bottleneck representations $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ as the distance metric to quantify their similarity. Given the set of cross-modal pairs $\mathcal{S} = \{s_{ij}\}$, the logarithm Maximum

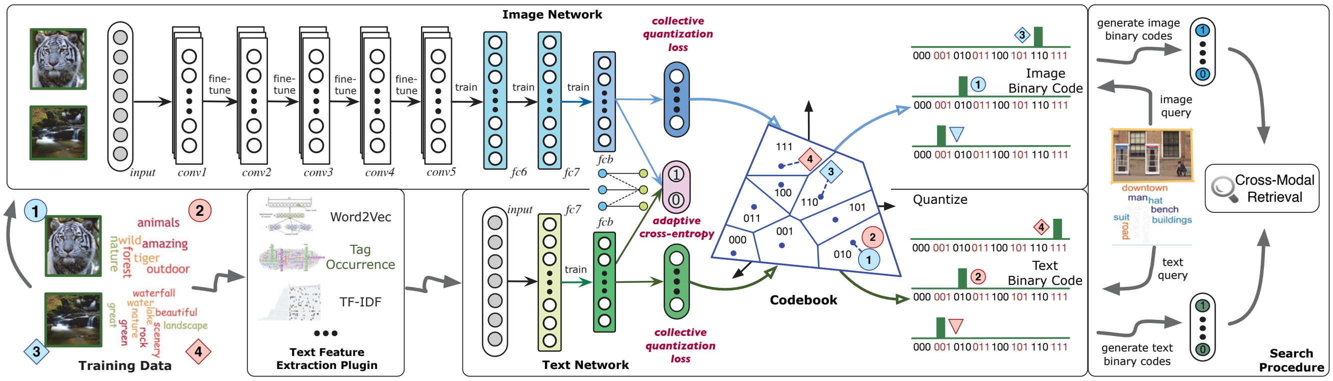


Figure 1: Collective Deep Quantization (CDQ) constitutes (1) a convolutional neural network (CNN) for learning image representations, (2) a multilayer perceptrons (MLP) for learning text representations, (3) two fully-connected bottleneck layers fcb for learning similarity-preserving representations, (4) an adaptive cross-entropy loss for capturing cross-modal correlations, and a collective quantization loss for controlling quantization error and enhancing the quantizability of bottleneck representation.

a Posteriori (MAP) estimation of bottleneck representations $\mathbf{Z}^x = [\mathbf{z}_1^x, \ldots, \mathbf{z}_{N_x}^x]$ and $\mathbf{Z}^y = [\mathbf{z}_1^y, \ldots, \mathbf{z}_{N_y}^y]$ is defined as

$$\begin{aligned}\log p\left(\mathbf{Z}^x, \mathbf{Z}^y | \mathcal{S}\right) &\propto \log p\left(\mathcal{S} | \mathbf{Z}^x, \mathbf{Z}^y\right) p\left(\mathbf{Z}^x\right) p\left(\mathbf{Z}^y\right) \\ &= \sum_{s_{ij} \in \mathcal{S}} \log p\left(s_{ij} | \mathbf{z}_i^x, \mathbf{z}_j^y\right) p\left(\mathbf{z}_i^x\right) p\left(\mathbf{z}_j^y\right),\end{aligned} \quad (1)$$

where $p(\mathcal{S}|\mathbf{Z}^x, \mathbf{Z}^y)$ is the likelihood function, $p(\mathbf{Z}^x)$ and $p(\mathbf{Z}^y)$ are prior distributions. For each pair of points $\mathbf{x}_i$ and $\mathbf{y}_j$, $p(s_{ij}|\mathbf{z}_i^x, \mathbf{z}_j^y)$ is the conditional probability of their relationship s_{ij} given their bottleneck representations $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$, which can be defined using the pairwise logistic function,

$$\begin{aligned}p\left(s_{ij} | \mathbf{z}_i^x, \mathbf{z}_j^y\right) &= \begin{cases} \sigma\left(\left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right), & s_{ij} = 1 \\ 1 - \sigma\left(\left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right), & s_{ij} = 0 \end{cases} \\ &= \sigma\left(\left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right)^{s_{ij}} \left(1 - \sigma\left(\left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right)\right)^{1 - s_{ij}},\end{aligned} \quad (2)$$

where $\sigma(x) = 1/(1 + e^{-\alpha x})$ is the *adaptive* sigmoid function with hyper-parameter α to control its bandwidth. Note that the sigmoid function with larger α will have larger saturation zone where its gradient is zero. To perform more effective back-propagation, we may require $\alpha < 1$, which is more effective than the typical setting of $\alpha = 1$. Similar to logistic regression, the larger the inner product $\langle \mathbf{z}_i^x, \mathbf{z}_j^y \rangle$ is, the larger $p(1|\mathbf{z}_i^x, \mathbf{z}_j^y)$ will be, implying that pair $\mathbf{x}_i$ and $\mathbf{y}_j$ should be classified as "similar"; otherwise, the larger $p(0|\mathbf{z}_i^x, \mathbf{z}_j^y)$ will be, implying that pair $\mathbf{x}_i$ and $\mathbf{y}_j$ should be classified as "dissimilar". Hence, Equation (2) is a reasonable extension of the logistic regression classifier to the pairwise classification scenario, which is optimal for binary pairwise labels $s_{ij} \in \{0, 1\}$. By MAP (2), the cross-modal correlation is maximized while the cross-modal relationship $\mathcal{S}$ can be preserved in the bottleneck representations.

It is worth noting that, to generate accurate binary codes from the bottleneck representations $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$, we need to enhance their *quantizability*—the possibility of being quantized to binary codes with infinitesimal error. We can improve the quantizability of bottleneck representations $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ such that they can be quantized more effectively by a specific quantizer, e.g. product quantization (Ge et al. 2014). As shown in (Ge et al. 2014), not all input vectors can be quantized effectively using the quantization technique—if input vectors do not exhibit a cluster/manifold structure, then they may not be quantized accurately, which is a common sense for data clustering. In this paper, we propose a novel Gaussian prior over the bottleneck representations $\mathbf{z}_i^x$ and $\mathbf{z}_j^y$ as

$$p\left(\mathbf{z}_i^*\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left\|\mathbf{z}_i^* - \hat{\mathbf{z}}_i^*\right\|^2 / 2\sigma^2\right), \quad (3)$$

where $* \in \{x, y\}$, σ and $\hat{\mathbf{z}}_i^*$ are the covariance and mean parameters of the Gaussian distribution. It is important to note that, $\hat{\mathbf{z}}_i^*$ is assumed to be a vector that can be perfectly quantized using existing quantization techniques, i.e. its quantization error can be zero. Thus, we can use a specific quantization technique to perfectly reconstruct $\hat{\mathbf{z}}_i^*$. We observe that maximizing this prior is reduced to minimizing the squared loss between the bottleneck representation $\hat{\mathbf{z}}_i$ and the corresponding perfectly quantizable vector $\hat{\mathbf{z}}_i^*$, which improves the quantizability of the bottleneck representations.

By substituting Equations (2) and (3) into the MAP estimation in Equation (1), we achieve the optimization problem to maximize cross-modal correlation and quantizability as

$$\min O = L + \lambda Q, \quad (4)$$

where $\lambda = 1/2\sigma^2$ is the trade-off parameter between adaptive cross-entropy loss L and collective quantization loss Q. Specifically, the adaptive cross-entropy loss L is defined as

$$L = \sum_{s_{ij} \in \mathcal{S}} \log\left(1 + \exp\left(\alpha \left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right)\right) - \alpha s_{ij} \left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle. \quad (5)$$

Note that α is the hyper-parameter of the adaptive sigmoid function that enables effective back-propagation for network training. Similarly, the quantization loss Q can be derived as

$$Q = N_y \sum\nolimits_{i=1}^{N_x} \left\|\mathbf{z}_i^x - \hat{\mathbf{z}}_i^x\right\|^2 + N_x \sum\nolimits_{j=1}^{N_y} \left\|\mathbf{z}_j^y - \hat{\mathbf{z}}_j^y\right\|^2. \quad (6)$$

At now, the perfectly quantizable representations $\hat{\mathbf{z}}_i^*$ are still unknown variables. We can use a specific quantization technique to perfectly reconstruct $\hat{\mathbf{z}}_i^*$ such that the quantization error is zero. We will adopt the state-of-the-art composite quantization (Zhang, Du, and Wang 2014) as the quantizer.

Collective Quantization

Based on composite quantization (Zhang, Du, and Wang 2014), the quantizable representations $\hat{\mathbf{z}}_i^*$ can be perfectly quantized by a set of M codebooks $\mathbf{C}^* = [\mathbf{C}_1^*, \ldots, \mathbf{C}_M^*]$, where each codebook $\mathbf{C}_m^*$ contains K codewords $\mathbf{C}_m^* = [\mathbf{C}_{m1}^*, \ldots, \mathbf{C}_{mK}^*]$, each codeword $\mathbf{C}_{mk}^*$ is a R-dimensional vector like the cluster centroid in kmeans clustering. Corresponding to the M codebooks, we partition the binary codewords assignment vector $\mathbf{b}_i^*$ into M 1-of-K indicator vectors $\mathbf{b}_i^* = [\mathbf{b}_{1i}^*; \ldots; \mathbf{b}_{mi}^*]$, and each indicator vector $\mathbf{b}_{mi}^*$ indicates which one (and only one) of the K codewords in the mth codebook is selected to approximate the ith data point. The composite quantizer encodes each $\mathbf{x}_i^*$ as the sum of M codewords, one codeword per codebook, each indicated by the binary assignment vector $\mathbf{b}_i^*$. This yields an error-free composite reconstruction $\hat{\mathbf{z}}_i^* = \sum_{m=1}^M \mathbf{C}_m^* \mathbf{b}_{mi}^*$. To enable knowledge sharing across different modalities such that cross-modal correlation can be further maximized, we propose a *collective quantization* approach by sharing the codebooks $\{\mathbf{C}_m^* = \mathbf{C}_m\}_{m=1}^M$ across different modalities, which can serve as a bridge in the common feature space for knowledge transfer. By substituting the composite reconstruction $\hat{\mathbf{z}}_i^* = \sum_{m=1}^M \mathbf{C}_m \mathbf{b}_{mi}^*$ and codebook sharing in Equation (6),

$$Q = N_y \sum_{i=1}^{N_x} \left\| \mathbf{z}_i^x - \sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi}^x \right\|^2 + N_x \sum_{j=1}^{N_y} \left\| \mathbf{z}_j^y - \sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mj}^y \right\|^2, \tag{7}$$

where $\|\mathbf{b}_{mi}^*\|_0 = 1, \mathbf{b}_{mi}^* \in \{0,1\}^K, * \in \{x, y\}$, $\|\cdot\|_0$ is the ℓ_0-norm that simply counts the number of the vector's nonzero elements. The constraint guarantees that only one codeword in each codebook can be activated to approximate the input data, which leads to compact binary codes. The rationale of using M codebooks instead of single codebook to approximate each input datum is to further minimize quantization error, as the latter is shown to yield significantly lossy compression and incur evident performance drop (Zhang, Du, and Wang 2014; Babenko and Lempitsky 2014).

Approximate nearest neighbor (ANN) search based on Inner Product distance is a powerful task for quantization techniques (Du and Wang 2014). Given an image database $\{\mathbf{b}_i^x\}_{i=1}^{N_x}$, we use *Asymmetric Quantizer Distance* (AQD) (Long et al. 2016) as similarity metric that computes the distance between text query $\mathbf{q}^y$ and image database point $\mathbf{x}_i$ as

$$\mathrm{AQD}\left(\mathbf{q}^y, \mathbf{x}_i\right) = \mathbf{z}_q^{y\mathsf{T}} \cdot \left(\sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi}^x \right), \tag{8}$$

Given a query, these inner products for all M codebooks $\{\mathbf{C}_m\}_{m=1}^M$ and all K possible values of $\mathbf{b}_{mi}^x$ can be precomputed and stored in a query-specific $M \times K$ lookup table, which is used to compute AQD between the query and all database points, each entails M table lookups and additions and is slightly more costly than Hamming distance.

Learning Algorithm

The CDQ optimization problem in Equation (4) consists of three sets of variables, network parameters Θ, shared codebook $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_M]$, and binary codes $\mathbf{B}^*$. We adopt an alternating optimization paradigm (Long et al. 2016) that iteratively updates one variable with the rest variables fixed.

The network parameters Θ can be efficiently optimized through standard back-propagation (BP) algorithm via the automatic differentiation techniques in Google TensorFlow.

We update the shared codebook $\mathbf{C}$ by fixing Θ and $\mathbf{B}$ as known variables, and write Equation (4) with $\mathbf{C}$ as unknown variables in matrix formulation,

$$\min_{\mathbf{C}} N_y \left\| \mathbf{Z}^x - \mathbf{C}\mathbf{B}^x \right\|^2 + N_x \left\| \mathbf{Z}^y - \mathbf{C}\mathbf{B}^y \right\|^2. \tag{9}$$

This is a quadratic problem with analytic solution $\mathbf{C} = \left[N_y \mathbf{Z}^x \mathbf{B}^{x\mathsf{T}} + N_x \mathbf{Z}^y \mathbf{B}^{y\mathsf{T}} \right] \left[N_y \mathbf{B}^x \mathbf{B}^{x\mathsf{T}} + N_x \mathbf{B}^y \mathbf{B}^{y\mathsf{T}} \right]^{-1}$. Algorithms such as L-BFGS can speed up the computation.

As each $\mathbf{b}_i^x$ is independent on $\{\mathbf{b}_{i'}^x\}_{i' \neq i}$, the optimization problem for $\mathbf{B}^x$ is decomposed to N_x subproblems,

$$\begin{aligned} &\min_{\mathbf{b}_i^x} \left\| \mathbf{z}_i^x - \sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi}^x \right\|^2 \\ &\text{s.t.} \quad \|\mathbf{b}_{mi}^x\|_0 = 1, \mathbf{b}_{mi}^x \in \{0,1\}^K. \end{aligned} \tag{10}$$

This optimization problem is generally NP-hard. As shown in (Zhang, Du, and Wang 2014), this problem is essentially high-order Markov Random Field (MRF) problem and can be solved by the Iterated Conditional Modes (ICM) algorithm (Besag 1986) which solves M indicators $\{\mathbf{b}_{mi}^x\}_{m=1}^M$ alternatively. Given $\{\mathbf{b}_{m'i}^x\}_{m' \neq m}$ fixed, we update $\mathbf{b}_{mi}^x$ by exhaustively checking all the codeword in codebook $\mathbf{C}_m$, finding the codeword such that the objective in (10) is minimized, and setting the corresponding entry of $\mathbf{b}_{mi}^x$ as 1 and the rest as 0. The ICM algorithm is guaranteed to converge, and can be terminated if maximum iterations are reached.

Approximation Error Analysis

Given a text query $\mathbf{q}^y$ and an image database point $\mathbf{x}_i$ with binary code $\mathbf{b}_i^x$, their inner product distance can be computed as $d\left(\mathbf{q}^y, \mathbf{x}_i\right) = \mathbf{z}_q^{y\mathsf{T}} \mathbf{z}_i^x$, where $\mathbf{z}_q^y$ and $\mathbf{z}_i^x$ are the deep representations of $\mathbf{q}^y$ and $\mathbf{x}_i$, respectively. We analyze approximation error of using AQD (8) to approximate the inner product distance. Denote by $\hat{\mathbf{z}}_i^x = \mathbf{C}\mathbf{b}_i^x$ the reconstruction of $\mathbf{u}_i^l$ using Equation (7), then AQD $\left(\mathbf{q}^y, \mathbf{x}_i\right) = d\left(\mathbf{q}^y, \hat{\mathbf{x}}_i\right)$.

Theorem 1 (Error Bound)**.** *The error of AQD* (8) *to approximate original inner-product distance is bounded by loss* (7)

$$\left| AQD\left(\mathbf{q}^y, \mathbf{x}_i\right) - d\left(\mathbf{q}^y, \mathbf{x}_i\right) \right| \leqslant \left\| \mathbf{z}_q^y \right\| \left\| \mathbf{z}_i^x - \mathbf{C}\mathbf{b}_i^x \right\|. \tag{11}$$

Proof. From the triangle inequality, it follows that

$$\begin{aligned} \left| \mathrm{AQD}\left(\mathbf{q}^y, \mathbf{x}_i\right) - d\left(\mathbf{q}^y, \mathbf{x}_i\right) \right| &= \left| d\left(\mathbf{q}^y, \hat{\mathbf{x}}_i\right) - d\left(\mathbf{q}^y, \mathbf{x}_i\right) \right| \\ &= \left| \mathbf{z}_q^{y\mathsf{T}} \left(\mathbf{z}_i^x - \mathbf{C}\mathbf{b}_i^x \right) \right| \\ &\leqslant \left\| \mathbf{z}_q^y \right\| \left\| \mathbf{z}_i^x - \mathbf{C}\mathbf{b}_i^x \right\|. \end{aligned}$$

□

Table 1: Mean Average Precision (MAP) Comparison of Two Cross-Modal Retrieval Tasks on Two Datasets

Task	Method	NUS-WIDE				MIR-Flickr			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
$I \to T$	CMSSH (Bronstein et al. 2010)	0.4535	0.4665	0.4752	0.4809	0.5062	0.5122	0.5247	0.5404
	CVH (Kumar and Udupa 2011)	0.4368	0.4454	0.4462	0.4342	0.6724	0.6883	0.6968	0.7092
	IMH (Song et al. 2013)	0.5042	0.5256	0.5629	0.6358	0.6698	0.6765	0.6825	0.6989
	SCM (Zhang and Li 2014)	0.6659	0.6871	0.7134	0.7271	0.6868	0.6953	0.7014	0.7091
	SePH (Lin et al. 2015)	0.5962	0.5982	0.6018	0.5910	0.7468	0.7526	0.7592	0.7604
	CCQ (Long et al. 2016)	0.6874	0.6879	0.7221	0.7347	0.6140	0.6052	0.6133	0.6656
	MMNN (Masci et al. 2014)	0.6147	0.6255	0.6296	0.6424	0.6825	0.6915	0.7024	0.7185
	CHN (Cao, Long, and Wang 2016)	0.7126	0.7692	0.8018	0.8220	0.7890	0.8430	0.8492	0.8571
	CDQ	**0.8227**	**0.8495**	**0.8488**	**0.8492**	**0.8479**	**0.8635**	**0.8602**	**0.8618**
$T \to I$	CMSSH (Bronstein et al. 2010)	0.4313	0.4166	0.4719	0.5110	0.4468	0.4656	0.4652	0.4624
	CVH (Kumar and Udupa 2011)	0.4148	0.4357	0.4419	0.4253	0.5918	0.6065	0.6148	0.6277
	IMH (Song et al. 2013)	0.6024	0.6253	0.6625	0.6816	0.6052	0.6229	0.6197	0.6201
	SCM (Zhang and Li 2014)	0.6528	0.6794	0.6972	0.7194	0.5972	0.6173	0.6192	0.6115
	SePH (Lin et al. 2015)	0.5969	0.6044	0.6011	0.6036	0.6259	0.6470	0.6479	0.6429
	CCQ (Long et al. 2016)	0.6182	0.6740	0.6940	0.6982	0.6378	0.6486	0.6499	0.6593
	MMNN (Masci et al. 2014)	0.5997	0.6083	0.6148	0.6226	0.6711	0.6815	0.6898	0.6992
	CHN (Cao, Long, and Wang 2016)	0.7170	0.7617	0.7689	0.7704	0.7595	0.7631	0.7725	0.7814
	CDQ	**0.8144**	**0.8321**	**0.8426**	**0.8478**	**0.8292**	**0.8477**	**0.8482**	**0.8495**

Since $\left\|\mathbf{z}_q^y\right\|$ does not affect the relative ordering of the inner product distances between $\mathbf{q}^y$ and all database points $\{\mathbf{x}_i\}_{i=1}^{N_x}$, minimizing the collective quantization loss (7) will give low error when performing approximate nearest neighbor search based on AQD instead of inner product distance.

Experiments

Setup

NUS-WIDE (Chua et al. 2009) is a public web image dataset. There are 81 ground truth concepts manually annotated for search evaluation. Following prior works (Wang et al. 2014b; Zhu et al. 2013), 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** (Huiskes and Lew 2008) consists of 25,000 images collected from the Flickr website, where each image is labeled with some of the 38 semantic concepts. We resize images of this labeled subset into 256×256.

For our deep learning based approach CDQ, we directly use the raw image pixels as the input. For fair comparison, for traditional shallow hashing methods, we use AlexNet (Krizhevsky, Sutskever, and Hinton 2012) to 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.

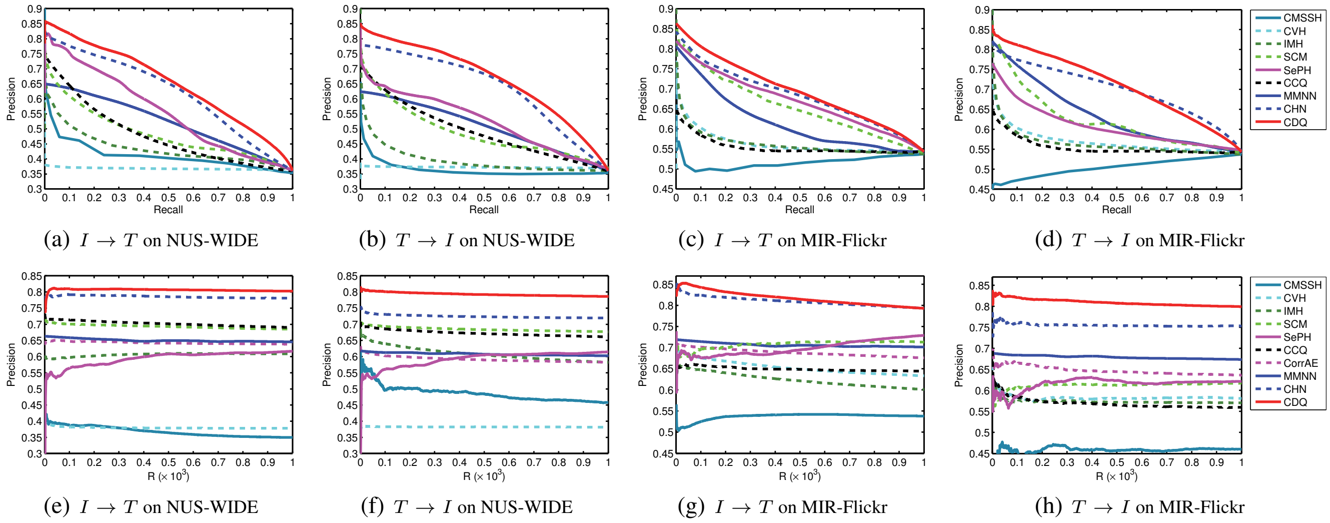
We compare CDQ with many state-of-the-art cross-modal hashing, quantization and deep hashing methods, including four unsupervised methods **IMH** (Song et al. 2013), **CVH** (Kumar and Udupa 2011), **CCQ** (Long et al. 2016) and **MMNN** (Masci et al. 2014), and four supervised methods **CMSSH** (Bronstein et al. 2010), **SCM** (Zhang and Li 2014), **SePH** (Lin et al. 2015) and **CHN** (Cao, Long, and Wang 2016), where **CCQ** is a cross-modal quantization method, **MMNN** and **CHN** are cross-modal deep hashing methods.

We follow protocols in (Cao, Long, and Wang 2016; Long et al. 2016) to evaluate the retrieval quality based on three metrics: Mean Average Precision (MAP) with MAP@$R = 50$, precision-recall curves and precision@top-R curves.

We implement the CDQ model via **TensorFlow**. For image network, we use AlexNet (Krizhevsky, Sutskever, and Hinton 2012), fine-tune $conv1$–$fc7$ copied from the pre-trained model, and train bottlneck layer fcb, all via back-propagation. For text network, we use a two-layer multilayer perceptrons (MLP), in which the $fc7$ layer has 4096 ReLU units with dropout rate 0.5, and the fcb layer have R $\tanh$ units. We use mini-batch SGD with 0.9 momentum, fix mini-batch size as 64, and cross-validate the learning rate. We follow similar strategy in (Long et al. 2016): (1) set the dimension of bottleneck layer $D = 128$ such that the composite quantizer can quantize the bottleneck representations accurately; (2) set $K = 256$ codewords for each codebook; (3) for each data point, the binary code of all M subspaces requires $B = M \log_2 K = 8M$ bits (i.e. M bytes) for compact coding, where we set $M = B/8$ as B is known. We select parameters of all methods via cross-validation. Each experiment repeats five runs and average results are reported.

Results

We report in Table 1 the MAP of all methods with different lengths of binary codes, i.e. 8, 16, 24 and 32 bits. We can observe that CDQ substantially outperforms all state-of-the-art methods for all cross-modal retrieval tasks. Specifically, compared to the best shallow baseline SCM, CDQ achieves absolute increases of 14.42% / 14.70%, 16.02% / 23.23% in average MAP for $I \to T$ / $T \to I$ on NUS-WIDE and MIR-Flickr. Compared to the deep cross-modal hashing methods,



(a) $I \to T$ on NUS-WIDE

(b) $T \to I$ on NUS-WIDE

(c) $I \to T$ on MIR-Flickr

(d) $T \to I$ on MIR-Flickr

(e) $I \to T$ on NUS-WIDE

(f) $T \to I$ on NUS-WIDE

(g) $I \to T$ on MIR-Flickr

(h) $T \to I$ on MIR-Flickr

Figure 2: Precision-recall curves (a)-(d) and Precision@top R curves (e)-(h) on NUS-WIDE and MIR-Flickr with 32 bits codes.

Table 2: Mean Average Precision (MAP) Comparison of CDQ Variants on NUS-WIDE and MIR-Flickr

Task	Method	NUS-WIDE				MIR-Flickr			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
$I \to T$	CDQ-ip	0.6987	0.7102	0.7157	0.7171	0.7359	0.7576	0.7625	0.7687
	CDQ-α	0.8052	0.8241	0.8332	0.8369	0.8315	0.8409	0.8436	0.8432
	CDQ-2	0.7861	0.8295	0.8306	0.8354	0.8158	0.8399	0.8425	0.8585
	CDQ	**0.8227**	**0.8495**	**0.8488**	**0.8492**	**0.8479**	**0.8635**	**0.8602**	**0.8618**
$T \to I$	CDQ-ip	0.6089	0.6254	0.6298	0.6363	0.7146	0.7265	0.7316	0.7391
	CDQ-α	0.7985	0.8201	0.8268	0.8325	0.8026	0.8187	0.8236	0.8267
	CDQ-2	0.8006	0.8131	0.8274	0.8302	0.8106	0.8249	0.8267	0.8376
	CDQ	**0.8144**	**0.8321**	**0.8426**	**0.8478**	**0.8292**	**0.8477**	**0.8482**	**0.8495**

CDQ outperforms state-of-the-art CHN by large margins of 6.61% / 7.97%, 2.38% / 7.45% in average MAP. These results verify that CDQ is able to learn high-quality binary codes for effective cross-modal retrieval.

We respectively report in Figure 2 (a)-(d) the precision-recall curves with 32 bits for two cross-modal retrieval tasks $I \to T$ and $T \to I$ on two benchmark datasets NUS-WIDE and MIR-Flickr. CDQ shows the best retrieval performance at all recall levels. Figure 2 (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. CDQ significantly outperforms all state-of-the-art methods under these metrics.

Discussion

To go deeper with the efficacy of CDQ, we design three variants of the proposed approach: a two-step method **CDQ-2**, which separately learns bottleneck representations via deep networks and the compositional binary codes via collective quantization loss (7); **CDQ-α** is the CDQ variant with $\alpha = 1$ in adaptive cross-entropy loss (5); **CDQ-ip** is the CDQ variant that utilizes the widely-adopted pairwise inner-product loss $L = \sum_{s_{ij} \in \mathcal{S}} \left(s_{ij} - \frac{1}{B} \left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle \right)^2$ (Liu et al. 2012; Xia et al. 2014) instead of the proposed adaptive cross-entropy loss (5). The MAP results are reported in Table 2.

From Table 2, we have the following key observations. **(a)** By simultaneously preserving similarity information using adaptive cross-entropy loss (5) and controlling hashing quality using collective quantization loss (7), CDQ outperforms CDQ-2 by 2.21% / 1.64%, 1.92% / 1.87% in average MAP. It is indispensable to improve the quantizability of deep representations by optimizing the collective quantization loss when training the deep networks. **(b)** By using the adaptive cross-entropy loss (5), CDQ can outperform CDQ-ip using the pairwise inner-product loss by very large margins of 13.21% / 20.91%, 10.22% / 11.57% in average MAP. The pairwise inner-product loss has been widely adopted in previous work (Liu et al. 2012; Xia et al. 2014). However, this loss does not link well the pairwise distances between points (taking values in $(-D, D)$ when using continuous relaxation) to the pairwise similarity labels (taking binary values {-1,1}). In contrast, the proposed adaptive cross-entropy loss is inherently consistent with the training pairs. **(c)** Also, by introducing adaptive coefficient α into cross-entropy loss, CDQ outperforms CDQ-α by 1.77% / 1.48%, 1.86% / 2.57% in average MAP. This validates that the adaptive cross-entropy loss with wider non-saturation zone can be trained more effectively. In summary, these experimental results also imply that all the components in CDQ are important for achieving the promising performance, and missing any component will result in substantial performance drop.

Conclusion

In this paper, we have proposed a novel Collective Deep Quantization (CDQ) for effective and efficient cross-modal retrieval. CDQ is a hybrid deep architecture that jointly optimizes the new adaptive cross-entropy loss on semantic similarity pairs and the novel collective quantization loss for compact binary codes. Extensive experiments on standard cross-modal retrieval datasets validate that CDQ can yield substantial boosts over state-of-the-art hashing methods.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61325008, 61502265), National Key R&D Program of China (2016YFB1000701, 2015BAF32B01), and Tsinghua National Laboratory (TNList) Key Project.

References

Babenko, A., and Lempitsky, V. 2014. Additive quantization for extreme vector compression. In *CVPR*. IEEE.

Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society* 48(3):259–320.

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

Cao, Y.; Long, M.; Wang, J.; Yang, Q.; and Yu, P. S. 2016a. Deep visual-semantic hashing for cross-modal retrieval. In *SIGKDD*.

Cao, Y.; Long, M.; Wang, J.; Zhu, H.; and Wen, Q. 2016b. Deep quantization network for efficient image retrieval. In *AAAI*.

Cao, Y.; Long, M.; Wang, J.; and Zhu, H. 2016c. Correlation autoencoder hashing for supervised cross-modal search. In *ICMR*. ACM.

Cao, Y.; Long, M.; and Wang, J. 2016. Correlation hashing network for efficient cross-modal retrieval. *CoRR* abs/1602.06697.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. 2009. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*. ACM.

Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Du, C., and Wang, J. 2014. Inner product similarity search using compositional codes. *CoRR* abs/1406.4966.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, 2121–2129.

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*.

Ge, T.; He, K.; Ke, Q.; and Sun, J. 2014. Optimized product quantization. *TPAMI*.

Hu, Y.; Jin, Z.; Ren, H.; Cai, D.; and He, X. 2014. Iterative multi-view hashing for cross media indexing. In *MM*. ACM.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *ICMR*. ACM.

Jiang, Q., and Li, W. 2016. Deep cross-modal hashing. *CoRR* abs/1602.02255.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *NIPS*.

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

Kumar, S., and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *IJCAI*.

Lai, H.; Pan, Y.; Liu, Y.; and Yan, S. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*.

Lin, Z.; Ding, G.; Hu, M.; and Wang, J. 2015. Semantics-preserving hashing for cross-view retrieval. In *CVPR*.

Lin, M.; Chen, Q.; and Yan, S. 2014. Network in network. In *ICLR*.

Liu, W.; Wang, J.; Ji, R.; Jiang, Y.-G.; and Chang, S.-F. 2012. Supervised hashing with kernels. In *CVPR*. IEEE.

Liu, X.; He, J.; Deng, C.; and Lang, B. 2014. Collaborative hashing. In *CVPR*. IEEE.

Long, M.; Cao, Y.; Wang, J.; and Yu, P. S. 2016. Composite correlation quantization for efficient multimodal retrieval. In *SIGIR*.

Masci, J.; Bronstein, M. M.; Bronstein, A. M.; and Schmidhuber, J. 2014. Multimodal similarity-preserving hashing. *TPAMI* 36.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. MIT Press. chapter Learning Internal Representations by Error Propagation.

Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; and Shen, H. T. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD*. ACM.

Wang, J.; Shen, H. T.; Song, J.; and Ji, J. 2014a. Hashing for similarity search: A survey. Arxiv.

Wang, W.; Ooi, B. C.; Yang, X.; Zhang, D.; and Zhuang, Y. 2014b. Effective multi-modal retrieval based on stacked auto-encoders. In *VLDB*. ACM.

Weiss, Y.; Torralba, A.; and Fergus, R. 2009. Spectral hashing. In *NIPS*.

Xia, R.; Pan, Y.; Lai, H.; Liu, C.; and Yan, S. 2014. Supervised hashing for image retrieval via image representation learning. In *AAAI*.

Yu, Z.; Wu, F.; Yang, Y.; Tian, Q.; Luo, J.; and Zhuang, Y. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *SIGIR*. ACM.

Zhang, D., and Li, W. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*.

Zhang, T., and Wang, J. 2016. Collaborative quantization for cross-modal similarity search. In *CVPR*.

Zhang, T.; Du, C.; and Wang, J. 2014. Composite quantization for approximate nearest neighbor search. In *ICML*. ACM.

Zhen, Y., and Yeung, D. 2012a. Co-regularized hashing for multimodal data. In *NIPS*, 1385–1393.

Zhen, Y., and Yeung, D.-Y. 2012b. A probabilistic model for multimodal hash function learning. In *SIGKDD*. ACM.

Zhu, X.; Huang, Z.; Shen, H. T.; and Zhao, X. 2013. Linear cross-modal hashing for efficient multimedia search. In *MM*. ACM.

Zhu, H.; Long, M.; Wang, J.; and Cao, Y. 2016. Deep hashing network for efficient similarity retrieval. In *AAAI*.