Transitive Hashing Network for Heterogeneous Multimedia Retrieval*

Zhangjie Cao†, Mingsheng Long†, Jianmin Wang†, Qiang Yang‡
†KLiss, MOE; TNList; School of Software, Tsinghua University, China
‡Hong Kong University of Science and Technology, Hong Kong
caozhangjie14@gmail.com {mingsheng,jimwang}@tsinghua.edu.cn qyang@cse.ust.hk

Abstract

Hashing is widely applied to large-scale multimedia retrieval due to the storage and retrieval efficiency. Cross-modal hashing enables efficient retrieval of one modality from database relevant to a query of another modality. Existing work on cross-modal hashing assumes that heterogeneous relationship across modalities is available for learning to hash. This paper relaxes this strict assumption by only requiring heterogeneous relationship in some auxiliary dataset different from the query or database domain. We design a novel hybrid deep architecture, transitive hashing network (THN), to jointly learn cross-modal correlation from the auxiliary dataset, and align the data distributions of the auxiliary dataset with that of the query or database domain, which generates compact transitive hash codes for efficient cross-modal retrieval. Comprehensive empirical evidence validates that the proposed THN approach yields state of the art retrieval performance on standard multimedia benchmarks, i.e. NUS-WIDE and ImageNet-YahooQA.

Introduction

Multimedia retrieval has attracted increasing attention in the presence of multimedia big data emerging in search engines and social networks. Cross-modal retrieval is an important paradigm of multimedia retrieval, which supports similarity retrieval across different modalities, e.g. retrieval of relevant images in response to text queries. A promising solution to cross-modal retrieval is hashing methods, which compress high-dimensional data into compact binary codes and generate similar codes for similar objects (Wang et al. 2014a). To date, however, effective and efficient cross-modal hashing remains a challenge, due to the heterogeneity across modalities (Wei et al. 2014), and the semantic gap between features and semantics (Smeulders et al. 2000).

An overview of cross-modal retrieval problems is shown in Figure 1. Prior cross-modal hashing methods (Bronstein et al. 2010; Kumar and Udupa 2011; Zhen and Yeung 2012; Song et al. 2013; Masci et al. 2014; Zhang and Li 2014; Wu et al. 2015; Jiang and Li 2016) have achieved promising performance for multimedia retrieval. However, they all require that the heterogeneous relationship between query and database is available for hash function learning. This is a rather strong requirement for many practical applications, where such heterogeneous relationship is not available. For example, a user of YahooQA (Yahoo Answers) may hope to search images relevant to his QAs from an online social media such as ImageNet. Unfortunately, because there are no link connections between YahooQA and ImageNet, it is not easy to satisfy the user's information need. Therefore, how to support cross-modal retrieval without direct heterogeneous relationship between query and database is an interesting open problem worth investigation.

This paper proposes a novel transitive hashing network (THN) approach to cross-modal retrieval without direct heterogeneous relationship between query and database, which generates compact hash codes of images and texts in an end-to-end deep learning architecture to construct the transitivity between query and database of different modalities. As learning cross-modal correlation is impossible without any heterogeneous relationship information, we leverage an auxiliary dataset readily available from a different but related domain (such as Flickr.com), which contains related heterogeneous relationship (e.g. images and their associated texts). We craft a hybrid deep network to enable heterogeneous relationship learning on this auxiliary dataset. As the auxiliary dataset and the query/database are collected from different domains and follow different data distributions, there is substantial dataset shift which poses a major difficulty to bridge them. To this end, we design and integrate a homogeneous distribution alignment module to the hybrid deep network, which closes the gap between the auxiliary dataset and the query/database. Based on heterogeneous relationship learning and homogeneous distribution alignment, we can construct the transitivity between query and database in an end-to-end deep architecture to enable efficient heterogeneous multimedia retrieval. Extensive experiments show that THN yields state of the art multimedia retrieval performance on public benchmarks NUS-WIDE and ImageNet-YahooQA.

Related Work

This work is related to hashing for multimedia retrieval, a.k.a. cross-modal hashing, which has been an increasingly popular research topic in the machine learning, computer vision, and multimedia retrieval communities (Bronstein et

*Corresponding authors: Mingsheng Long and Jianmin Wang.

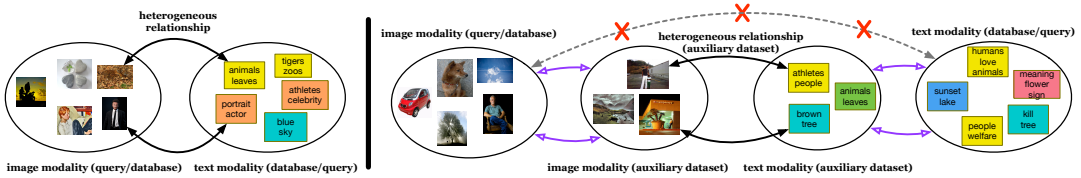


Figure 1: Problem overview. (left) Prior cross-modal hashing, where heterogeneous relationship between query and database (black arrows) is available for hash learning. (right) Our transitive hashing, where heterogeneous relationship is not available between query and database (dashed arrows) but is available from an auxiliary dataset of different distributions (purple arrows).

al. 2010; Kumar and Udupa 2011; Zhen and Yeung 2012; Song et al. 2013; Masci et al. 2014; Zhang and Li 2014; Wu et al. 2015; Jiang and Li 2016; Cao et al. 2016). We refer readers to (Wang et al. 2014a) for a comprehensive survey.

Previous cross-modal hashing methods can be organized into unsupervised methods and supervised methods. Unsupervised methods learn hash functions that convert input data points into binary codes only using unlabeled data. Typical learning criteria include reconstruction error minimization (Wang et al. 2014b), neighborhood preserving in graph-based hashing (Kumar and Udupa 2011; Song et al. 2013), and quantization error minimization in correlation quantization (Wu et al. 2015; Long et al. 2016). Supervised methods explore supervised information (e.g. pairwise similarity or relevance feedback) to learn more discriminative compact hash codes. Typical learning criteria include metric learning (Bronstein et al. 2010), neural network (Masci et al. 2014), and correlation learning (Zhang and Li 2014; Wu et al. 2015). As supervised methods explore the semantic relationship to bridge modalities and reduce the semantic gap (Smeulders et al. 2000), they can achieve superior accuracy than unsupervised methods for cross-modal retrieval.

Prior cross-modal hashing methods based on shallow architectures cannot effectively exploit the heterogeneous relationship across different modalities. Latest work on deep multimodal embedding(Frome et al. 2013; Kiros, Salakhutdinov, and Zemel 2014; Donahue et al. 2015; Gao et al. 2015) has shown that deep models can bridge heterogeneous modalities more effectively for image description and understanding, but it remains unclear how to explore these deep models for cross-modal hashing. Recent deep hashing methods (Xia et al. 2014; Lai et al. 2015; Zhu et al. 2016) have given state of the art results on many image datasets, but they can only be used for single-modal retrieval. To the best of our knowledge, Deep Cross-Modal Hashing (DCMH) (Jiang and Li 2016) and Correlation Hashing Network (CHN) (Cao, Long, and Wang 2016) are 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, DCMH and CHN can only address traditional cross-modal retrieval where heterogeneous relationship between query and database is available for hash learning, which is very restricted for real applications. The proposed transitive hashing network (THN) addresses cross-modal retrieval where heterogeneous relationship is not available between query and database, which leverages an auxiliary cross-modal dataset in another domain and builds transitivity to bridge query and database.

Transitive Hashing Network

In the transitive hashing problem, we are given a query set $\mathcal{X}^q = \{\boldsymbol{x}_i\}_{i=1}^n$ from modality X, and a database set $\mathcal{Y}^d = \{\boldsymbol{y}_j\}_{j=1}^m$ from modality Y, where $\boldsymbol{x}_i \in \mathcal{R}^{d_x}$ is a d_x-dimensional feature vector in the query modality and $\boldsymbol{y}_j \in \mathcal{R}^{d_y}$ is a d_y-dimensional feature vector in the database modality. The key challenge of transitive hashing is that no supervised relationship is available between query and database. Hence, we bridge modalities X and Y by learning from an auxiliary dataset $\bar{\mathcal{X}} = \{\bar{\boldsymbol{x}}_i\}_{i=1}^{\bar{n}}$ and $\bar{\mathcal{Y}} = \{\bar{\boldsymbol{y}}_j\}_{j=1}^{\bar{m}}$ available in a different domain, which comprises cross-modal relationship $\mathcal{S} = \{s_{ij}\}$, where $s_{ij} = 1$ implies points $\bar{\boldsymbol{x}}_i$ and $\bar{\boldsymbol{y}}_j$ are similar while $s_{ij} = 0$ indicates points $\bar{\boldsymbol{x}}_i$ and $\bar{\boldsymbol{y}}_j$ are dissimilar. In real multimedia retrieval applications, the cross-modal relationship $\mathcal{S} = \{s_{ij}\}$ can be collected from the relevance feedbacks in click-through data, or from the social media where multiple modalities are available.

The goal of transitive hashing network (THN) is to learn two hash functions $f_x : \mathbb{R}^{d_x} \to \{-1, 1\}^b$ and $f_y : \mathbb{R}^{d_y} \to \{-1, 1\}^b$ that encode data points from modalities X and Y into compact b-bit hash codes $\boldsymbol{h}_x = f_x(\boldsymbol{x})$ and $\boldsymbol{h}_y = f_y(\boldsymbol{y})$ respectively, such that the cross-modal relationship $\mathcal{S}$ can be preserved. With the learned hash functions, we can generate hash codes $\mathcal{H}^q = \{\boldsymbol{h}_i^x\}_{i=1}^n$ and $\mathcal{H}^d = \{\boldsymbol{h}_j^y\}_{j=1}^m$ for the query modality and database modality respectively, which enables multimedia retrieval across heterogeneous data based on ranking the Hamming distances between hash codes.

We learn transitive hash functions f_x and f_y by constructing the training sets $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ and $\mathcal{Y} = \{\boldsymbol{y}_j\}_{j=1}^M$ as follows: (1) $\mathcal{X}$ comprises the whole auxiliary dataset $\bar{\mathcal{X}}$ and another $\hat{n}$ data points randomly selected from the query set $\mathcal{X}^q$, where $N = \bar{n} + \hat{n}$; (2) $\mathcal{Y}$ comprises the whole auxiliary dataset $\bar{\mathcal{Y}}$ and another $\hat{m}$ data points randomly selected from the database set $\mathcal{Y}^d$, where $M = \bar{m} + \hat{m}$.

Architecture for Transitive Hashing

The architecture for learning transitive hash functions is shown in Figure 2, which is a hybrid deep architecture of 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

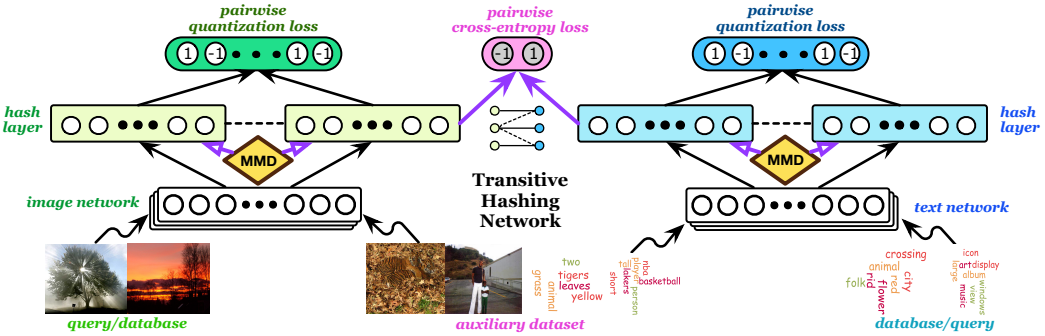


Figure 2: Transitive hashing network (THN), which comprises heterogeneous relationship learning, homogeneous distribution alignment, quantization error minimization, builds a transitivity (in purple) from query to database across modalities/domains.

fch hash layer with b hidden units, which transforms the network activation $\boldsymbol{z}_i^x$ in b-bit hash code by sign thresholding $\boldsymbol{h}_i^x = \mathrm{sgn}(\boldsymbol{z}_i^x)$. 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 fch hash layer with b hidden units to transform the network activation $\boldsymbol{z}_i^y$ in b-bit hash code by sign thresholding $\boldsymbol{h}_i^y = \mathrm{sgn}(\boldsymbol{z}_i^y)$. We adopt the hyperbolic tangent (tanh) function to squash the activations to be within $[-1, 1]$, which reduces the gap between the fch-layer representation $\boldsymbol{z}_i^*$ and the binary hash codes $\boldsymbol{h}_i^*$, where $* \in \{x, y\}$. Several carefully-designed loss functions on the hash codes are added on top of the hybrid deep network for heterogeneous relationship learning and homogeneous distribution alignment, which enable query-database transitivity construction for heterogeneous multimedia retrieval.

Heterogeneous Relationship Learning

In this work, we jointly preserve the heterogeneous relationship $\mathcal{S}$ in Hamming space and control the quantization error of sign thresholding in a Bayesian framework. We bridge the Hamming spaces of modalities X and Y by learning from the auxiliary dataset $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$. Note that, for a pair of binary codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, there exists a nice relationship between their Hamming distance $\mathrm{dist}_H(\cdot,\cdot)$ and their inner product $\langle \cdot, \cdot \rangle$: $\mathrm{dist}_H\left(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \frac{1}{2}\left(K - \left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right)$. Hence we will use inner product as a good surrogate of Hamming distance to quantify similarity between hash codes. Given heterogeneous relationship $\mathcal{S} = \{s_{ij}\}$, the logarithm Maximum a Posteriori (MAP) estimation of hash codes $\boldsymbol{H}^x = [\boldsymbol{h}_1^x, \ldots, \boldsymbol{h}_{\bar{n}}^x]$ and $\boldsymbol{H}^y = [\boldsymbol{h}_1^y, \ldots, \boldsymbol{h}_{\bar{m}}^y]$ can be defined as

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

where $p(\mathcal{S}|\boldsymbol{H}^x, \boldsymbol{H}^y)$ is likelihood function, and $p(\boldsymbol{H}^x)$ and $p(\boldsymbol{H}^y)$ are prior distributions. For each pair of points $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, $p(s_{ij}|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ is the conditional probability of their relationship s_{ij} given their hash codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, which can be defined using the pairwise logistic function as follows,

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

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function and $\boldsymbol{h}_i^x = \mathrm{sgn}(\boldsymbol{z}_i^x)$ and $\boldsymbol{h}_j^y = \mathrm{sgn}(\boldsymbol{z}_j^y)$. Similar to logistic regression, the smaller the Hamming distance $\mathrm{dist}_H(\boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ is, the larger the inner product $\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \rangle$ will be, and the larger $p(1|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ will be, implying that pair $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$ should be classified as "similar"; otherwise, the larger $p(0|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ will be, implying that pair $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$ 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 heterogeneous relationship $\mathcal{S}$ can be preserved in the Hamming space.

Since discrete optimization of Equation (1) with binary constraints $\boldsymbol{h}_i^* \in \{-1, 1\}^b$ is difficult, for ease of optimization, continuous relaxation that $\boldsymbol{h}_i^x = \boldsymbol{z}_i^x$ and $\boldsymbol{h}_j^y = \boldsymbol{z}_j^y$ is applied to the binary constraints, as widely adopted by existing hashing methods (Wang et al. 2014a). To reduce the gap between the binary hash codes and continuous network activations, We adopt the hyperbolic tangent (tanh) function to squash the activations to be within $[-1, 1]$. However, the continuous relaxation still gives rise to two issues: (1) uncontrollable quantization error by binarizing continuous activations to binary codes, and (2) large approximation error by adopting inner product between continuous activations as the surrogate of Hamming distance between binary codes. In this paper, to control the quantization error and close the gap between Hamming distance and its surrogate for learning accurate hash codes, we propose a new cross-entropy prior over the continuous activations $\{\boldsymbol{z}_i^*\}$ as

$$p\left(\boldsymbol{z}_i^*\right) \propto \exp\left(-\lambda H\left(\frac{\mathbf{1}}{b}, \frac{|\boldsymbol{z}_i^*|}{b}\right)\right), \tag{3}$$

where $* \in \{x, y\}$, and λ is the parameter of the exponential distribution. We observe that maximizing this prior is reduced to minimizing the cross-entropy $H(\cdot,\cdot)$ between the

uniform distribution $\mathbf{1}/b$ and the code distribution $|\boldsymbol{z}_i^*|/b$, which is equivalent to assigning each bit of the continuous activations $\{\boldsymbol{z}_i^*\}$ to binary values $\{-1, 1\}$.

By substituting Equations (2) and (3) into the MAP estimation in Equation (1), we achieve the optimization problem for heterogeneous relationship learning as follows,

$$\min_{\Theta} J = L + \lambda Q, \tag{4}$$

where λ is the trade-off between pairwise cross-entropy loss L and pairwise quantization loss Q, and Θ denotes the set of network parameters. Specifically, loss L is defined as

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

Similarly the pairwise quantization loss Q can be derived as

$$Q = \sum_{s_{ij} \in \mathcal{S}} \sum_{k=1}^{b} (-\log(|z_{ik}^x|) - \log(|z_{jk}^y|)). \tag{6}$$

By the MAP estimation in Equation (4), we can simultaneously preserve the heterogeneous relationship in training data and control the quantization error of binarizing continuous activations to binary codes. By learning from the auxiliary dataset, we can successfully bridge different modalities.

Homogeneous Distribution Alignment

The goal of transitive hashing is to perform efficient retrieval from the database of one modality in response to the query of another modality. Since there is no relationship between the query and the database, we exploit the auxiliary dataset $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$ to bridge the query modality and database modality. However, since the auxiliary dataset is obtained from a different domain, there are large distribution shifts between the auxiliary dataset and the query/database sets. Therefore, we should further reduce the distribution shifts by minimizing the Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) between the auxiliary dataset and the query set (or between the auxiliary dataset and the database set) in the Hamming space. MMD is a nonparametric distance measure to compare different distributions P_d and P_x in reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) endowed with feature map ϕ and kernel k (Gretton et al. 2012), formally defined as $D_q \triangleq \left\| \mathbb{E}_{\boldsymbol{h}^q \sim P_q} \left[\phi\left(\boldsymbol{h}^q\right)\right] - \mathbb{E}_{\boldsymbol{h}^x \sim P_x} \left[\phi\left(\boldsymbol{h}^x\right)\right] \right\|_{\mathcal{H}}^2$, where P_q is the distribution of the query set $\mathcal{X}^q$, and P_x is the distribution of the auxiliary set $\bar{\mathcal{X}}$. Using the same continuous relaxation as previous section, the MMD between the auxiliary dataset $\bar{\mathcal{X}}$ and the query set $\mathcal{X}^q$ can be computed as

$$\begin{aligned} D_q = &\sum_{i=1}^{\hat{n}} \sum_{j=1}^{\hat{n}} \frac{k\left(\boldsymbol{z}_i^q, \boldsymbol{z}_j^q\right)}{\hat{n}^2} + \sum_{i=1}^{\bar{n}} \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^x\right)}{\bar{n}^2} \\ &- 2 \sum_{i=1}^{\hat{n}} \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^q, \boldsymbol{z}_j^x\right)}{\hat{n}\bar{n}}, \end{aligned} \tag{7}$$

where $k(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-\gamma||\boldsymbol{z}_i - \boldsymbol{z}_j||^2)$ is the Gaussian kernel. Similarly, the MMD D_d between the auxiliary dataset $\bar{\mathcal{Y}}$ and the query set $\mathcal{Y}^d$ can be computed by replacing the query modality with the database modality, i.e. by replacing q, x, $\hat{n}$, $\bar{n}$ with d, y, $\hat{m}$, $\bar{m}$ in Equation (7), respectively.

Transitive Hash Function Learning

To enable efficient retrieval from the database of one modality in response to the query of another modality, we construct the transitivity bridge between the query and the database (as shown by the purple arrows in Figure 2) by integrating the objective functions of heterogeneous relationship learning (4) and the homogeneous distribution alignment (7) into a unified optimization problem as

$$\min_{\Theta} C = J + \mu \left(D_q + D_d\right), \tag{8}$$

where μ is a trade-off parameter between the MAP loss J and the MMD penalty $(D_q + D_d)$. By optimizing the objective function in Equation (8), we can learn transitive hash codes which preserve the heterogeneous relationship and align the homogeneous distributions as well as control the quantization error of sign thresholding. Finally, we generate b-bit hash codes by sign thresholding as $\boldsymbol{h}^* = \operatorname{sgn}(\boldsymbol{z}^*)$, where $\operatorname{sgn}(\boldsymbol{z})$ is the sign function on vectors that for each dimension i of $\boldsymbol{z}^*$, $i = 1, 2, ..., b$, $\operatorname{sgn}(z_i^*) = 1$ if $z_i^* > 0$, otherwise $\operatorname{sgn}(z_i^*) = -1$. Since the quantization error in Equation (8) has been minimized, this final binarization step will incur small loss of retrieval quality.

We derive the learning algorithms for the proposed transitive hashing network (THN) model in Equation (8) through the standard back-propagation (BP) algorithm. For clarity, we denote the point-wise cost with respect to $\bar{\boldsymbol{x}}_i$ as

$$\begin{aligned} C_i = &\sum_{j: s_{ij} \in \mathcal{S}} \log\left(1 + \exp\left(\left\langle \boldsymbol{z}_i^x, \boldsymbol{z}_j^y \right\rangle\right)\right) - s_{ij} \left\langle \boldsymbol{z}_i^x, \boldsymbol{z}_j^y \right\rangle \\ &- \lambda \sum_{j: s_{ij} \in \mathcal{S}} \sum_{k=1}^{b} \log(|z_{ik}^x|) + \mu \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^x\right)}{\bar{n}^2} - 2\mu \sum_{j=1}^{\hat{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^q\right)}{\hat{n}\bar{n}}. \end{aligned} \tag{9}$$

In order to run the BP algorithm, we only need to compute the residual term $\frac{\partial C_i}{\partial \tilde{z}_{ik}}$ for each data point $\bar{\boldsymbol{x}}_i$, where $\tilde{z}_{ik}^x$ is the output of the last layer before its activation function $a(\cdot) = \tanh(\cdot)$. We can derive the residual term as

$$\begin{aligned} \frac{\partial C_i}{\partial \tilde{z}_{ik}^x} = &\sum_{j: s_{ij} \in \mathcal{S}} \left(\left[\sigma\left(\left\langle \boldsymbol{z}_i^x, \boldsymbol{z}_j^y \right\rangle\right) - s_{ij}\right] z_{jk}^y \right) a'\left(\tilde{z}_{ik}^x\right) \\ &- \frac{\lambda}{z_{ik}^x} \sum_{j: s_{ij} \in \mathcal{S}} a'\left(\tilde{z}_{ik}^x\right) \\ &- 2\mu\gamma \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^x\right)}{\bar{n}^2} \left(z_{ik}^x - z_{jk}^x\right) a'\left(\tilde{z}_{ik}^x\right) \\ &+ 4\mu\gamma \sum_{j=1}^{\hat{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^q\right)}{\hat{n}\bar{n}} \left(z_{ik}^x - z_{jk}^q\right) a'\left(\tilde{z}_{ik}^x\right). \end{aligned} \tag{10}$$

The other residual terms with respect to modality Y can be derived similarly. Since the only difference between standard BP and our algorithm is Equation (10), we analyze the computational complexity based on Equation (10). Denote by $|\mathcal{S}|$ the number of relationship pairs $\mathcal{S}$ available for training, then it is easy to see that the computational complexity of the BP is $O(|\mathcal{S}| + BN)$, where B is the mini-batch size.

Experiments

Setup

NUS-WIDE[1] is a popular dataset for cross-modal retrieval, which contains 269,648 image-text pairs. The annotation for 81 semantic categories is provided for evaluation, which we prune by keeping the image-text pairs that belong to the 16 categories shared with ImageNet (Deng et al. 2009). Each image is resized into 256×256 pixels, and each text is represented by a bag-of-word (BoW) feature vector. We perform two types of cross-modal retrieval on the NUS-WIDE dataset: (1) using image query to retrieve texts (denoted by $I \rightarrow T$); (2) using text query to retrieve images (denoted by $T \rightarrow I$). The heterogeneous relationship $\mathcal{S}$ for training and the ground-truth for evaluation are defined as follows: if an image i and a text j (not necessarily from the same pair) share at least one of the 16 categories, they are relevant, i.e. $s_{ij} = 1$; otherwise, they are irrelevant, i.e. $s_{ij} = 0$.

ImageNet-YahooQA (Wei et al. 2014) is a heterogenous media dataset of images from ImageNet (Deng et al. 2009) and QAs from Yahoo Answers (YahooQA). ImageNet is an image database of over 1 million images. We select the images that belong to the 16 categories shared with the NUS-WIDE dataset. YahooQA is a text dataset of about 300,000 QAs crawled by Yahoo Query Language (YQL). Each QA is regarded as a text document and represented by bag-of-word (BoW) features. As the QAs are unlabeled, to enable evaluation, we assign one of the 16 category labels to each QA by checking whether the corresponding class words match that QA as in (Wei et al. 2014). Note that, though the selected datasets from NUS-WIDE and ImageNet/YahooQA share the same set of labels, their data distributions are significantly different as they are collected from different domains. We perform two types of cross-modal retrieval on the ImageNet-YahooQA dataset: (1) using image query in ImageNet to retrieve texts from YahooQA (denoted by $I \rightarrow T$); (2) using text query in YahooQA to retrieve images from ImageNet (denoted by $T \rightarrow I$). The ground-truth for evaluation is consistent with that of the NUS-WIDE dataset.

We follow (Wei et al. 2014) to evaluate the retrieval quality based on standard evaluation metrics: Mean Average Precision (MAP) and Precision-Recall curves. We compare the retrieval quality of our **THN** with five state of the art cross-modal hashing methods, including two unsupervised methods Cross-View Hashing (**CVH**) (Kumar and Udupa 2011) and Inter-Media Hashing (**IMH**) (Song et al. 2013), two supervised methods Quantized Correlation Hashing (**QCH**) (Wu et al. 2015) and Heterogeneous Translated Hashing (**HTH**) (Wei et al. 2014), and one deep hashing method Deep Cross-Modal Hashing (**DCMH**) (Jiang and Li 2016). Different from our method, QCH controls quantization error by Iterative Quantization (ITQ) (Gong and Lazebnik 2011).

For fair comparison, all of the methods use identical training and test sets. For deep learning based methods, including DCMH and the proposed THN, we directly use the image pixels as input. For the shallow learning based methods, we reduce the 4096-dimensional AlexNet features (Donahue et al. 2014) of images to 500 dimensions using PCA, which incurs negligible loss of retrieval quality but significantly speeds up the evaluation. For all methods, we use bag-of-word (BoW) features as text representations, which are reduced to 1000 dimensions by PCA to speed up evaluation.

We implement the THN model in **Caffe**. For image network, we adopt AlexNet (Krizhevsky, Sutskever, and Hinton 2012), fine-tune convolutional layer $conv1$–$conv5$ and fully-connected layer $fc6$–$fc7$ copied from the pre-trained model and train the fch hash layer from scratch, all via back-propagation. Since the hash layer fch is trained from scratch, we set its learning rate to be 10 times that of the other layers. For text network, we employ a three-layer MLP with the numbers of hidden units set to 1000, 500, and b, respectively. We use the mini-batch stochastic gradient descent (SGD) with 0.9 momentum and the learning rate strategy in Caffe, cross-validate learning rate from 10^{-5} to 10^{-1} with a multiplicative step-size $10^{1/2}$. We train the image network and the text network jointly in the hybrid deep architecture by optimizing the objective function in Equation (8). The codes and configurations will be made available online.

Results

NUS-WIDE: We follow the experimental protocols in (Wei et al. 2014). We randomly select 2,000 images or texts as query set, and correspondingly, the remaining texts and images are used as the database. We randomly select 30 images and 30 texts per class distinctly from the database as the training set, which means that the images and texts are not paired so the relationship between them are heterogeneous.

We compare the retrieval accuracies of the proposed THN with five state of the art hashing methods. The MAP results are presented in Table 1. We can observe that THN generally outperforms the comparison methods on the two cross-modal tasks. In particular, compared to the state of the art deep hashing method DCMH, we achieve relative increases of **9.47%** and **2.85%** in average MAP for the two cross-modal retrieval tasks $I \rightarrow T$ and $T \rightarrow I$ respectively.

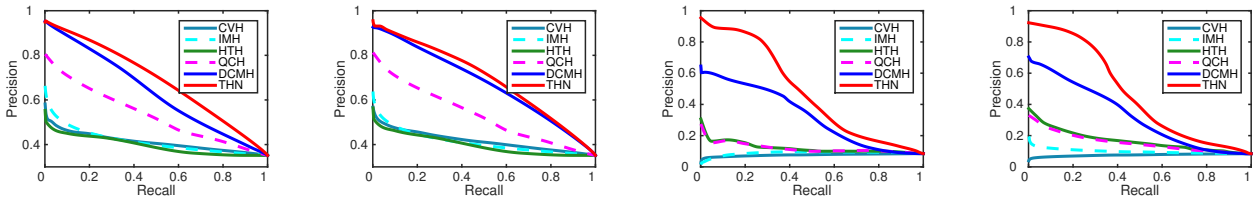
The precision-recall curves based on 24-bits hash codes for the two cross-modal retrieval tasks are illustrated in Figure 3. We can observe that THN achieves the highest precision at all recall levels. This results validate that THN is robust under diverse retrieval scenarios preferring either high precision or recall. The superior results in both MAP and precision-recall curves suggest that THN is a new state of the art method for the more conventional cross-modal retrieval problems where the relationship between query and database is available for training as in the NUS-WIDE dataset.

ImageNet-YahooQA: We follow similar protocols as in (Wei et al. 2014). We randomly select 2,000 images from ImageNet or 2000 texts from YahooQA as query set, and correspondingly, the remaining texts in YahooQA and the images in ImageNet are used as database. For the training set, we randomly select 2000 NUS-WIDE images and 2000 NUS-WIDE texts as supervised auxiliary dataset and select 500 ImageNet images and 500 Yahoo text documents as unsupervised training data. For all comparison methods, we note that they can only use the heterogeneous relationship in the supervised auxiliary dataset (NUS-WIDE) but cannot use the unsupervised training data from the query set and the

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Table 1: MAP Comparison of Cross-Modal Retrieval Tasks on NUS-WIDE and ImageNet-YahooQA Datasets

Task	Method	NUS-WIDE				ImageNet-YahooQA			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
$I \rightarrow T$	IMH (Song et al. 2013)	0.5821	0.5794	0.5804	0.5776	0.0855	0.0686	0.0999	0.0889
	CVH (Kumar and Udupa 2011)	0.5681	0.5606	0.5451	0.5558	0.1229	0.1180	0.0941	0.0865
	QCH (Wu et al. 2015)	0.6463	0.6921	0.7019	0.7127	0.2563	0.2494	0.2581	0.2590
	HTH (Wei et al. 2014)	0.5232	0.5548	0.5684	0.5325	0.2931	0.2694	0.2847	0.2663
	DCMH (Jiang and Li 2016)	0.7887	0.7397	0.7210	0.7460	0.5133	0.5109	0.5321	0.5087
	THN (ours)	**0.8252**	**0.8423**	**0.8495**	**0.8572**	**0.5451**	**0.5507**	**0.5803**	**0.5901**
$T \rightarrow I$	IMH (Song et al. 2013)	0.5579	0.5593	0.5528	0.5457	0.1105	0.1044	0.1183	0.0909
	CVH (Kumar and Udupa 2011)	0.5261	0.5193	0.5097	0.5045	0.0711	0.0728	0.1116	0.1008
	QCH (Wu et al. 2015)	0.6235	0.6609	0.6685	0.6773	0.2761	0.2847	0.2795	0.2665
	HTH (Wei et al. 2014)	0.5603	0.5910	0.5798	0.5812	0.2172	0.1702	0.3122	0.2873
	DCMH (Jiang and Li 2016)	0.7882	0.7912	0.7921	0.7718	0.5163	0.5510	0.5581	0.5444
	THN (ours)	**0.7905**	**0.8137**	**0.8245**	**0.8268**	**0.6032**	**0.6097**	**0.6232**	**0.6102**



(a) $I \rightarrow T$ @ 24 bits (b) $T \rightarrow I$ @ 24 bits (c) $I \rightarrow T$ @ 24 bits (d) $T \rightarrow I$ @ 24 bits

Figure 3: Precision-recall curves of Hamming ranking @ 24-bits codes on NUS-WIDE (a)-(b) and ImageNet-YahooQA (c)-(d).

database set (ImageNet and YahooQA). It is desirable that THN can use both supervised auxiliary dataset and unsupervised training data for heterogeneous multimedia retrieval.

The MAP results of all methods are compared in Table 1. We can observe that for these novel cross-modal and cross-domain retrieval tasks between ImageNet and YahooQA, THN outperforms the comparison methods on the two cross-modal tasks by very large margins. In particular, compared to state of the art deep hashing method DCMH, we achieve relative increases of **5.03%** and **6.91%** in average MAP for the cross-modal retrieval tasks $I \rightarrow T$ and $T \rightarrow I$ respectively, which is very impressive. Similarly, the precision-recall curves based on 24-bits hash codes for the two cross-modal and cross-domain retrieval tasks in Figure 3 show that THN achieves the highest precision at all recall levels.

The superior results of MAP and precision-recall curves suggest that THN is a powerful approach to learning transitive hash codes, which enables heterogeneous multimedia retrieval between query and database across both modalities and domains. THN integrates heterogeneous relationship learning, homogeneous distribution alignment, and quantization error minimization into an end-to-end hybrid deep architecture to build the transitivity between query and database. The results on the NUS-WIDE dataset already show that the heterogeneous relationship learning module is effective to bridge different modalities. The experiment on the ImageNet-YahooQA dataset further validates that the homogeneous distribution alignment between the auxiliary dataset and the query/database set, which is missing in all comparison methods, contributes significantly to the retrieval performance of THN. The reason is that the auxiliary dataset and the query/database sets are collected from different domains and follow different data distributions, hence there is substantial dataset shift which poses a major difficulty to bridge them. The homogeneous distribution alignment module of THN effectively closes this shift by matching the corresponding data distributions with the maximum mean discrepancy. This makes the proposed THN model a good fit to heterogeneous multimedia retrieval problems.

Discussion

We investigate the variants of THN on ImageNet-YahooQA: (1) **THN-ip** is the variant using the pairwise inner-product loss instead of the pairwise cross-entropy loss; (2) **THN-D** is the variant without using the unsupervised training data; (3) **THN-Q** is the variant without using the quantization loss. We report the MAP of all THN variants in Table 2. (1) THN outperforms THN-ip by very large margins of 24.15% / 23.19% in average MAP for cross-modal tasks $I \rightarrow T$ / $T \rightarrow I$, which confirms the importance of well-defined loss functions for heterogeneous relationship learning. (2) THN outperforms THN-D by 4.06% / 6.09% in average MAP. This validates that THN can further exploit the unsupervised training data to bridge the Hamming spaces of auxiliary dataset (NUS-WIDE) and query/database sets (ImageNet-YahooQA) such that the auxiliary dataset can transfer knowledge between query and database. (3) THN outperforms THN-Q by 5.83% / 4.20% in average MAP, which confirms that the quantization loss can reduce the errors of binarizing continuous representations to hash codes.

Table 2: MAP of THN variants on ImageNet-YahooQA

Method	$I \rightarrow T$				$T \rightarrow I$			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
THN-ip	0.2976	0.3171	0.3302	0.3554	0.3443	0.3605	0.3852	0.4286
THN-D	0.5192	0.5123	0.5312	0.5411	0.5423	0.5512	0.5602	0.5489
THN-Q	0.4821	0.5213	0.5352	0.4947	0.5731	0.5592	0.5849	0.5612
THN	**0.5451**	**0.5507**	**0.5803**	**0.5901**	**0.6032**	**0.6097**	**0.6232**	**0.6102**

Conclusion

In this paper, we have formally defined a new transitive deep hashing problem for heterogeneous multimedia retrieval, and proposed a novel transitive hashing network based on a hybrid deep architecture. The key to this problem is building the transitivity across different modalities and different data distributions, which relies on relationship learning and distribution alignment. Extensive empirical evidence on public multimedia datasets shows the proposed approach yields state of the art multimedia retrieval performance. In the future, we plan to extend the method to social media problems.

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

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. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *KDD*.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*.

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*.

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*.

Gong, Y., and Lazebnik, S. 2011. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*. IEEE.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *JMLR* 13:723–773.

Jiang, Q.-Y., and Li, W.-J. 2016. Deep cross-modal hashing. *arXiv preprint arXiv: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*. IEEE.

Long, M.; Cao, Y.; Wang, J.; and Yu, P. S. 2016. Composite correlation quantization for efficient multimodal retrieval. In *SIGIR*, SIGIR '16. ACM.

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.

Smeulders, A. W.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *TPAMI* 22.

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 preprint arXiv:1408.2927*.

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.

Wei, Y.; Song, Y.; Zhen, Y.; Liu, B.; and Yang, Q. 2014. Scalable heterogeneous translated hashing. In *KDD*, 791–800. ACM.

Wu, B.; Yang, Q.; Zheng, W.; Wang, Y.; and Wang, J. 2015. Quantized correlation hashing for fast cross-modal search. In *IJCAI*.

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

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

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

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