Transitive Hashing Network for Heterogeneous Multimedia Retrieval

Zhangjie Cao[†], Mingsheng Long[†], and Qiang Yang[‡]
[†]School of Software, Tsinghua University, Beijing 100084, China
[‡]Hong Kong University of Science and Technology, Hong Kong
caozhangjie14@gmail.com, mingsheng@tsinghua.edu.cn, qyang@cse.ust.hk

Abstract

Hashing has been widely applied to large-scale multimedia retrieval due to the storage and retrieval efficiency. Cross-modal hashing enables efficient retrieval from database of one modality in response to a query of another modality. Existing work on cross-modal hashing assumes heterogeneous relationship across modalities for hash function learning. In this paper, we relax the strong assumption by only requiring such heterogeneous relationship in an auxiliary dataset different from the query/database domain. We craft a hybrid deep architecture to simultaneously learn the cross-modal correlation from the auxiliary dataset, and align the dataset distributions between the auxiliary dataset and the query/database domain, which generates transitive hash codes for heterogeneous multimedia retrieval. Extensive experiments exhibit that the proposed approach yields state of the art multimedia retrieval performance on public datasets, i.e. NUS-WIDE, ImageNet-YahooQA.

1 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 with 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 [1]. To date, effective and efficient cross-modal hashing remains a challenge, due to the heterogeneity across modalities [2], and the semantic gap between features and semantics [3].

An overview of cross-modal retrieval problems is shown in Figure 1. Traditional cross-modal hashing methods [4, 5, 6, 7, 8, 9, 10, 11] 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 very strong requirement for many practical applications, where such heterogeneous relationship is not available. For example, a user of YahooQA (Yahoo Questions and Answers) may hope to search images relevant to his QAs from an online social media such as ImageNet. Unfortunately, since 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 relationship between query and database is an interesting problem worth investigation.

This paper proposes a novel transitive hashing network (THN) approach to address the above problem, 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 the heterogeneous relationship (e.g. images and their associated texts). We craft a hybrid deep network to enable heterogeneous relationship learning on this auxiliary dataset. Note that, the
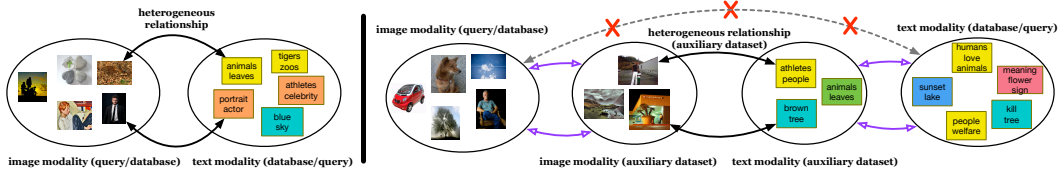
Figure 1: Problem overview. (left) Traditional cross-modal hashing, where heterogeneous relationship between query and database (black arrows) is available for hash learning. (right) The new transitive hashing, where heterogeneous relationship is not directly available between query and database (dashed arrows) but is available from an auxiliary dataset of different distributions (purple arrows).

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. To this end, we further integrate a homogeneous distribution alignment module to the hybrid deep network, which closes the gap between the auxiliary dataset and the query/database sets. 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 our THN model yields state of the art multimedia retrieval performance on public datasets, i.e. NUS-WIDE, ImageNet-YahooQA.

2 Related Work

This work is related to hashing for multimedia retrieval, known as cross-modal hashing, which has been an increasingly popular research topic in machine learning, computer vision, and multimedia retrieval communities [4, 5, 6, 7, 8, 9, 10, 11, 12]. Please refer to [1] for a comprehensive survey.

Previous cross-modal hashing methods can be organized into unsupervised methods and supervised methods. Unsupervised methods learn hash functions that encode input data points to binary codes only using unlabeled training data. Typical learning criteria include reconstruction error minimization [13], neighborhood preserving in graph-based hashing [5, 7], and quantization error minimization in correlation quantization [10, 14]. Supervised methods explore supervised information (e.g. pairwise similarity or relevance feedback) to learn compact hash codes. Typical learning criteria include metric learning [4], neural network [8], and correlation learning [9, 10]. As supervised methods can explore the semantic relationship to bridge modalities and reduce the semantic gap [3], they can achieve superior accuracy than unsupervised methods for cross-modal retrieval.

Most of previous cross-modal hashing methods based on shallow architectures cannot effectively exploit the heterogeneous relationship across different modalities. Latest deep models for multimodal embedding [15, 16, 17, 18] have shown that deep learning can bridge heterogeneous modalities more effectively for image captioning, but it remains unclear how to explore these deep models to cross-modal hashing. Recent deep hashing methods [19, 20, 21] have given state of the art results on many datasets, but they can only be used for single-modal retrieval. To the best of our knowledge, DCMH [11] is the only cross-modal deep hashing method that uses deep convolutional networks [22] for image representation and multilayer perceptrons [23] for text representation. However, DCMH 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. To this end, we propose a novel transitive hashing network (THN) method to address cross-modal retrieval where heterogeneous relationship is not available between query and database, which leverages an auxiliary cross-modal dataset from a different domain and builds transitivity to bridge different modalities.

3 Transitive Hashing Network

In transitive hashing, we are given a query set $\mathcal{X}^q = \{\mathbf{x}_i\}_{i=1}^n$ from modality X (such as image), and a database set $\mathcal{Y}^d = \{\mathbf{y}_j\}_{j=1}^m$ from modality Y (such as text), where $\mathbf{x}_i \in \mathbf{R}^{d_x}$ is a d_x-dimensional feature vector in the query modality and $\mathbf{y}_j \in \mathbf{R}^{d_y}$ is a d_y-dimensional feature vector in the database modality. A key challenge of transitive hashing is that no supervised relationship is available between query and database. Therefore, we bridge modalities X and Y by learning from an auxiliary dataset
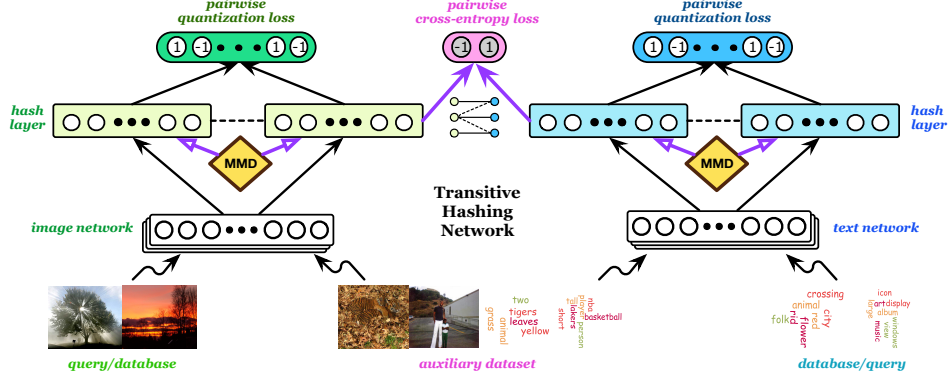
Figure 2: The hybrid architecture of Transitive Hashing Network (THN), which comprises heterogeneous relationship learning, homogeneous distribution alignment, and quantization error minimization. The key is to build a transitivity (in purple) from query to database across both modalities and domains.

$\bar{\mathcal{X}} = \{\bar{\mathbf{x}}_i\}_{i=1}^{\bar{n}}$ and $\bar{\mathcal{Y}} = \{\bar{\mathbf{y}}_j\}_{j=1}^{\bar{m}}$ available in a different domain, which comprises cross-modal relationship $\mathcal{S} = \{s_{ij}\}$, where $s_{ij} = 1$ implies points $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{y}}_i$ are similar while $s_{ij} = 0$ indicates points $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{y}}_i$ are disimilar. In real multimedia retrieval applications, the cross-modal relationship $S = \{s_{ij}\}$ can be collected from the relevance feedback information in click-through data, or from the social media where multiple modalities are usually presented.

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 $\mathbf{h}_x = f_x(\mathbf{x})$ and $\mathbf{h}_y = f_y(\mathbf{y})$ respectively, such that the cross-modal relationship $\mathcal{S}$ can be preserved. With the learned hash functions, we can generate the hash codes $\mathcal{H}^q = \{\mathbf{h}_i^x\}_{i=1}^n$ and $\mathcal{H}^d = \{\mathbf{h}_j^y\}_{j=1}^m$ for the query modality and database modality respectively, which enables multimedia retrieval across heterogeneous data based on ranking the Hamming distances between hash codes.

To learn the transitive hash functions f_x and f_y, we construct the training sets $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^M$ 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 $M = \bar{m} + \hat{m}$.

3.1 Architecture for Transitive Hashing

The architecture for learning transitive hash functions are illustrated in Figure 2, which is a hybrid deep architecture comprising an image network and a text network. In the image network, we extend AlexNet [22], 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 $\mathbf{z}_i^x$ in b-bit hash code by sign thresholding $\mathbf{h}_i^x = \text{sgn}(\mathbf{z}_i^x)$. In text network, we adopt the Multilayer perceptrons (MLP) [23] 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 $\mathbf{z}_i^y$ in b-bit hash code by sign thresholding $\mathbf{h}_i^y = \text{sgn}(\mathbf{z}_i^y)$. 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 $\mathbf{z}_i^*$ and the binary hash codes $\mathbf{h}_i^*$, where $* \in \{x, y\}$. Several carefully-designed loss functions on the hash codes are added on top of the hybrid network for heterogeneous relationship learning and homogeneous distribution alignment, which enables query-database transitivity construction for heterogeneous multimedia retrieval.

3.2 Heterogeneous Relationship Learning

In this work, we jointly preserve the heterogeneous relationship $\mathcal{S}$ in the 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 $\mathbf{h}_i^x$ and $\mathbf{h}_j^y$, there exists a nice linear relationship between their Hamming distance $\text{dist}_H(\cdot, \cdot)$

and inner product $\langle\cdot,\cdot\rangle$: $\mathrm{dist}_H\left(\mathbf{h}_i^x,\mathbf{h}_j^y\right)=\frac{1}{2}\left(K-\left\langle\mathbf{h}_i^x,\mathbf{h}_j^y\right\rangle\right)$. Hence in the sequel, we will use the inner product as a good surrogate of the Hamming distance to quantify the similarity between hash codes. Given heterogeneous relationship $\mathcal{S}=\{s_{ij}\}$, the logarithm Maximum a Posteriori (MAP) estimation of hash codes $\mathbf{H}^x=[\mathbf{h}_1^x,\ldots,\mathbf{h}_n^x]$ and $\mathbf{H}^y=[\mathbf{h}_1^y,\ldots,\mathbf{h}_m^y]$ can be defined as follows,

$$\log p\left(\mathbf{H^x},\mathbf{H^y}|\mathcal{S}\right)\propto \log p\left(\mathcal{S}|\mathbf{H^x},\mathbf{H^y}\right)p\left(\mathbf{H}^x\right)p\left(\mathbf{H}^y\right)$$

$$=\sum_{s_{ij}\in\mathcal{S}}\log p\left(s_{ij}|\mathbf{h}_i^x,\mathbf{h}_j^y\right)p\left(\mathbf{h}_i^x\right)p\left(\mathbf{h}_j^y\right), \tag{1}$$

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

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

$$=\sigma\left(\left\langle s_{ij}|\mathbf{h}_i^x,\mathbf{h}_j^y\right\rangle\right)^{s_{ij}}\left(1-\sigma\left(\left\langle s_{ij}|\mathbf{h}_i^x,\mathbf{h}_j^y\right\rangle\right)\right)^{1-s_{ij}},$$

where $\sigma\left(x\right)=1/(1+e^{-x})$ is the sigmoid function and note that $\mathbf{h}_i^x=\mathrm{sgn}(\mathbf{z}_i^x)$ and $\mathbf{h}_i^y=\mathrm{sgn}(\mathbf{z}_i^y)$. Similar to logistic regression, the smaller the Hamming distance $\mathrm{dist}_H(\mathbf{h}_i^x,\mathbf{h}_j^y)$ is, the larger the inner product $\langle\mathbf{h}_i^x,\mathbf{h}_j^y\rangle$ will be, and the larger $p(1|\mathbf{h}_i^x,\mathbf{h}_j^y)$ will be, implying that pair $\mathbf{h}_i^x$ and $\mathbf{h}_j^y$ should be classified as "similar"; otherwise, the larger $p(0|\mathbf{h}_i^x,\mathbf{h}_j^y)$ will be, implying that pair $\mathbf{h}_i^x$ and $\mathbf{h}_j^y$ 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 $\mathbf{h}_i^*\in\{-1,1\}^b$ is difficult, for ease of optimization, continuous relaxation that $\mathbf{h}_i^x=\mathbf{z}_i^x$ and $\mathbf{h}_i^y=\mathbf{z}_i^y$ is applied to the binary constraints, as widely adopted by existing hashing methods [1]. 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 $\{\mathbf{z}_i^*\}$, which is defined as follows,

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

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 $|\mathbf{z}_i^*|/b$, which is equivalent to assigning each bit of the continuous activations $\{\mathbf{z}_i^*\}$ 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 trade-off parameter between the pairwise cross-entropy loss L and the pairwise quantization loss Q, and Θ denotes the set of network parameters. Specifically, the pairwise cross-entropy loss L is defined as

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

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 optimizing 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.

3.3 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) [24] 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 [24], formally defined as $D_q \triangleq \left\| \mathbb{E}_{\mathbf{h}^q \sim P_q} \left[\phi \left(\mathbf{h}^q \right) \right] - \mathbb{E}_{\mathbf{h}^x \sim P_x} \left[\phi \left(\mathbf{h}^x \right) \right] \right\|_{\mathcal{H}}^2$, 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, the MMD between the auxiliary dataset $\bar{\mathcal{X}}$ and the query set $\mathcal{X}^q$ can be computed as

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

where $k(\mathbf{z_i}, \mathbf{z_j}) = \exp(-\gamma \|\mathbf{z_i} - \mathbf{z_j}\|^2)$ 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 $q, x, \hat{n}$ and $\bar{n}$ with $d, y, \hat{m}$, and $\bar{m}$ in Equation (7), respectively.

3.4 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 $\mathbf{h}^* = \text{sgn}(\mathbf{z}^*)$, where $\text{sgn}(\mathbf{z})$ is the sign function on vectors that for each dimension i of $\mathbf{z}^*$, $i = 1, 2, ..., b$, $\text{sgn}(z_i^*) = 1$ if $z_i^* > 0$, otherwise $\text{sgn}(z_i^*) = -1$. 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 THN model in Equation (8) through the standard back-propagation (BP) algorithm. For clarity, we denote the point-wise cost with respect to $\bar{\mathbf{x}}_i$ as

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

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

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

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 the number of relationship pairs $\mathcal{S}$ available for training as $|\mathcal{S}|$, then it is easy to verify that the computational complexity is $O(|\mathcal{S}| + BN)$, where B is mini-batch size.

4 Experiments

4.1 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 [25]. 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 $I \to T$); (2) using text query to retrieve images (denoted by $T \to I$). 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. relationship $s_{ij} = 1$; otherwise, they are irrelevant, i.e. relationship $s_{ij} = 0$.

ImageNet-YahooQA [2] is a heterogenous media dataset consisting of images from ImageNet [25] and QAs from Yahoo Questions and Answers[2] (YahooQA). ImageNet is an image database of over 1 million images organized according to the WordNet hierarchy. 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 from a public API of Yahoo Query Language (YQL), detailed in [2]. Each QA is regarded as a text document and represented by a bag-of-word (BoW) feature vector. 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 word matches that QA. Note that, though the selected datasets from NUS-WIDE and ImageNet/YahooQA share the same set of labels, their data distributions are significantly different since 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 $I \to T$); (2) using text query in YahooQA to retrieve images from ImageNet (denoted by $T \to I$). The ground-truth for evaluation is consistent with that of the NUS-WIDE dataset.

We follow [2] to evaluate the retrieval quality based on standard evaluation metrics: Mean Average Precision (MAP) and Precision-Recall curves. We evaluate and compare the retrieval quality of the proposed **THN** approach with five state of the art cross-modal hashing methods, including two unsupervised methods Cross-View Hashing (**CVH**) [5] and Inter-Media Hashing (**IMH**) [7], two supervised methods Quantized Correlation Hashing (**QCH**) [10] and Heterogeneous Translated Hashing (**HTH**) [2], and one deep hashing method Deep Cross-Modal Hashing (**DCMH**) [11].

For fair comparison, all of the methods use identical training and test sets. For the 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 [26] of images to 500 dimensions using PCA, which incurs negligible loss of retrieval quality but significantly speeds up the evaluation process. For all methods, we use bag-of-word (BoW) features for text representations, which are reduced to 1000 dimensions using PCA for speeding up the evaluation.

We implement the THN model in **Caffe**. For image network, we adopt AlexNet [22], 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 fch hash layer 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.

4.2 Results

NUS-WIDE: We follow the experimental protocols in [2]. 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.

---

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[2] http://developer.yahoo.com/yql/

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

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 [7]	0.5821	0.5794	0.5804	0.5776	0.0855	0.0686	0.0999	0.0889
	CVH [5]	0.5681	0.5606	0.5451	0.5558	0.1229	0.1180	0.0941	0.0865
	QCH [10]	0.6463	0.6921	0.7019	0.7127	0.2563	0.2494	0.2581	0.2590
	HTH [2]	0.5232	0.5548	0.5684	0.5325	0.2931	0.2694	0.2847	0.2663
	DCMH [11]	0.7887	0.7397	0.7210	0.7460	0.5133	0.5109	0.5321	0.5087
	THN	**0.8252**	**0.8423**	**0.8495**	**0.8572**	**0.5451**	**0.5507**	**0.5803**	**0.5901**
$T \rightarrow I$	IMH [7]	0.5579	0.5593	0.5528	0.5457	0.1105	0.1044	0.1183	0.0909
	CVH [5]	0.5261	0.5193	0.5097	0.5045	0.0711	0.0728	0.1116	0.1008
	QCH [10]	0.6235	0.6609	0.6685	0.6773	0.2761	0.2847	0.2795	0.2665
	HTH [2]	0.5603	0.5910	0.5798	0.5812	0.2172	0.1702	0.3122	0.2873
	DCMH [11]	0.7882	0.7912	0.7921	0.7718	0.5163	0.5510	0.5581	0.5444
	THN	**0.7905**	**0.8137**	**0.8245**	**0.8268**	**0.6032**	**0.6097**	**0.6232**	**0.6102**

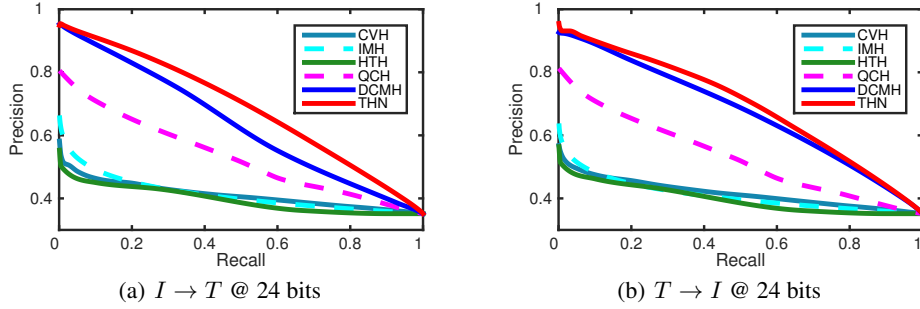(a) $I \rightarrow T$ @ 24 bits

(b) $T \rightarrow I$ @ 24 bits

Figure 3: Precision-recall curves of Hamming ranking with 24-bits hash codes on NUS-WIDE.

We evaluate and 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 in [2]. 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 the database. For the training set, we randomly select 2000 NUS-WIDE images and 2000 NUS-WIDE texts as the 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 database set (ImageNet and YahooQA). It is desirable that the THN model can use both supervised auxiliary dataset and unsupervised training data for heterogeneous multimedia retrieval.

We evaluate and 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 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 the state of the art deep hashing method DCMH, we achieve relative increases of **5.03%** and **6.91%** in average MAP for the two cross-modal retrieval tasks $I \rightarrow T$ and $T \rightarrow I$ respectively. Similarly, the precision-recall curves based on 24-bits hash codes for the two cross-modal and cross-domain retrieval tasks in Figure 4 shows that THN achieves the highest precision at all recall levels.

The superior results in both MAP and precision-recall curves suggest that THN is a powerful approach to learning transitive hash codes, which enables heterogeneous multimedia retrieval between query

Table 2: MAP Comparison of Cross-Modal Retrieval Tasks 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	<u>0.5192</u>	0.5123	0.5312	<u>0.5411</u>	0.5423	0.5512	0.5602	0.5489
THN-Q	0.4821	<u>0.5213</u>	<u>0.5352</u>	0.4947	<u>0.5731</u>	<u>0.5592</u>	<u>0.5849</u>	<u>0.5612</u>
THN	**0.5451**	**0.5507**	**0.5803**	**0.5901**	**0.6032**	**0.6097**	**0.6232**	**0.6102**

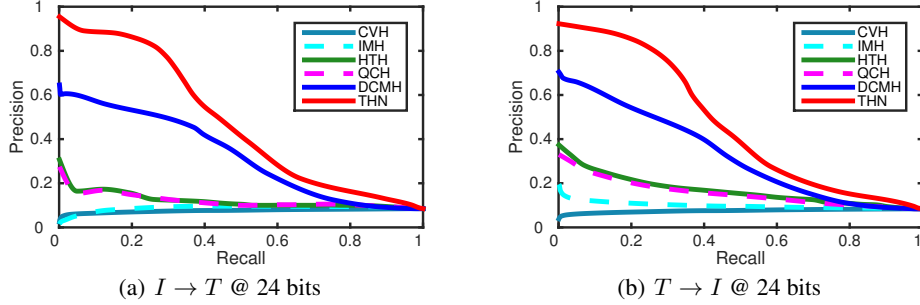(a) $I \rightarrow T$ @ 24 bits   (b) $T \rightarrow I$ @ 24 bits

Figure 4: Precision-recall curves of Hamming ranking with 24-bits hash codes on Imagenet-YahooQA.

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 for inferring the transitivity between query and database. The results on the NUS-WIDE dataset already shows 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 close 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.

4.3 Discussion

In order to study the effectiveness of THN, we investigate its variants on the ImageNet-YahooQA dataset: (1) **THN-ip** is the variant which uses 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 pairwise quantization loss. We report the MAP of all THN variants on ImageNet-Yahoo in Table 2. We may have the following observations. (1) THN outperforms THN-ip by very large margins of 24.15% / 23.19% in absolute increase of average MAP, which confirms the importance of well-defined loss functions for heterogeneous relationship learning. (2) Compared to THN-D, THN achieves absolute increases of 4.06% / 6.09% in average MAP for the two cross-modal tasks $I \rightarrow T$ and $T \rightarrow I$. This convinces 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 be served as a bridge to transfer knowledge between query and database. (3) THN also outperforms THN-Q by absolute promotions of 5.83% / 4.20% in average MAP, which confirms that the pairwise quantization loss can evidently reduce the quantization errors when binarizing the continuous representations to hash codes.

5 Conclusion

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

References

[1] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *arXiv preprint arXiv:1408.2927*, 2014.

[2] Y. Wei, Y. Song, Y. Zhen, B. Liu, and Q. Yang, "Scalable heterogeneous translated hashing," in *KDD*. ACM, 2014, pp. 791–800.

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

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

[5] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011.

[6] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *NIPS*, 2012.

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

[8] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *TPAMI*, vol. 36, 2014.

[9] D. Zhang and W. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *AAAI*, 2014.

[10] B. Wu, Q. Yang, W. Zheng, Y. Wang, and J. Wang, "Quantized correlation hashing for fast cross-modal search," in *IJCAI*, 2015.

[11] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," *arXiv preprint arXiv:1602.02255*, 2016.

[12] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *KDD*, 2016.

[13] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," in *VLDB*. ACM, 2014.

[14] M. Long, Y. Cao, J. Wang, and P. S. Yu, "Composite correlation quantization for efficient multimodal retrieval," in *SIGIR*, ser. SIGIR '16. ACM, 2016.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[16] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *NIPS*, 2014.

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

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

[19] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *AAAI*. AAAI, 2014.

[20] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *CVPR*. IEEE, 2015.

[21] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *AAAI 2016*, 2015.

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

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

[24] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *JMLR*, vol. 13, pp. 723–773, Mar. 2012.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

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