# Correlation Autoencoder Hashing for Supervised Cross-Modal Search*

Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu
School of Software, Tsinghua University, Beijing, China
Tsinghua National Laboratory for Information Science and Technology
{caoyue10, zhuhan10}@gmail.com {mingsheng, jimwang}@tsinghua.edu.cn

## ABSTRACT

Due to its storage and query efficiency, hashing has been widely applied to approximate nearest neighbor search from large-scale datasets. While there is increasing interest in cross-modal hashing which facilitates cross-media retrieval by embedding data from different modalities into a common Hamming space, how to distill the cross-modal correlation structure effectively remains a challenging problem. In this paper, we propose a novel supervised cross-modal hashing method, Correlation Autoencoder Hashing (CAH), to learn discriminative and compact binary codes based on deep autoencoders. Specifically, CAH jointly maximizes the feature correlation revealed by bimodal data and the semantic correlation conveyed in similarity labels, while embeds them into hash codes by nonlinear deep autoencoders. Extensive experiments clearly show the superior effectiveness and efficiency of CAH against the state-of-the-art hashing methods on standard cross-modal retrieval benchmarks.

## 1. INTRODUCTION

While big data with large volume, high dimensions, and multiple modalities are pervasive in search engines and social networks, it has attracted increasing attention to distill the correlation across heterogenous data modalities. For instance, an uploaded image on Flickr may be annotated with some relevant descriptions or tags, while a featured article on Wikipedia may consist of some correlative images. As relevant data from different modalities may endow semantic correlations, it is desirable to support *cross-modal search* that retrieves semantically-relevant results of one modality in response to a query of different modality. Taking Flickr as example, when a query image is given, the system should return not only relevant images but also relevant tags. Due to large volume and the well-known semantic gap [19], effective and efficient cross-modal retrieval remains a challenge.

When the reference database is large-scale or that the distance calculation between query item and database item is costly, an efficient solution to enabling similarity search is hashing methods [27], which perform approximate nearest neighbor (ANN) search with computation efficiency and search quality. The principle of hashing is to transform high-dimensional data into binary codes and generate similar binary codes for similar data items. The seminal work includes Locality Sensitive Hashing (LSH) [1], Spectral Hashing (SH) [29] and Product Quantization (PQ) [9]. However, the *unimodal* hashing methods cannot enable cross-modal search because ANN cannot be computed across modalities.

Recently, several *cross-modal hashing* methods have been proposed in the literature, which constructs correlation structures across modalities in the process of hash function learning and indexes cross-modal data into an isomorphic Hamming space [3, 12, 33, 34, 20, 28, 30, 14, 31, 6, 15]. These methods can be categorized into unsupervised methods [12, 33, 20] and supervised methods [3, 14, 31]. While unsupervised methods are more general and can be trained without semantic labels, they are also limited by the scarcity of correlation information and the semantic gap issue [19]. Supervised methods, on the contrary, can well explore the semantic labels for enhancing the cross-modal correlations and reducing the semantic gap [19], hence they generally outperform unsupervised methods for cross-modal search. The latest cross-modal retrieval models via deep learning [21, 28, 6, 24, 4] have shown that deep models can distill complex cross-modal correlations more effectively. Despite of the success of deep models for cross-modal search, existing cross-modal retrieval methods are mainly unsupervised and not tailored to hash function learning. Hence it remains an open problem how to explore both feature correlation revealed by bimodal data, and semantic correlation conveyed in similarity labels, using a unified deep model for cross-modal hashing.

In this paper, we propose Correlation Autoencoder Hashing (CAH), a novel model towards supervised cross-modal hashing via deep learning. CAH embeds data from different modalities into a common Hamming space by jointly maximizing the feature correlation revealed by bimodal data and the semantic correlation conveyed in similarity labels. More specifically: (1) We explore the *feature correlation* by reconstructing the feature vectors of one modality from the corresponding hash codes of another modality, and capture the cross-modal correlations revealed by the feature vectors; (2) We explore the *semantic correlation* by maximizing the inter-category separation margin and minimizing the intra-category variance, which will produce more discriminative and semantically consistent hash codes; (3) Since cross-modal data (e.g. image and text) are heterogeneous and are difficult to correlate by linear or shallow models [28, 6, 4], we enhance both cross-modal correlations in a deep architecture, which will make the embedded hash codes generalize better across different modalities. Comprehensive results on large-scale benchmarks show that CAH significantly outperforms state-of-the-art cross-modal hashing methods.

The rest of this paper is organized as follows. We review re-

*Corresponding author: Mingsheng Long.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*ICMR'16, June 06-09, 2016, New York, NY, USA*

© 2016 ACM. ISBN 978-1-4503-4359-6/16/06...$15.00

DOI: http://dx.doi.org/10.1145/2911996.2912000

lated works in Section 2. We formally present our correlation autoencoder hashing method in Section 3. Empirical evaluation is reported in Section 4, while the conclusion is enclosed in Section 5.

## 2. RELATED WORK

Recently, hashing-based cross-modal search has been a prevalent research focus in machine learning, computer vision, and multimedia search communities [3, 6, 7, 12, 14, 20, 25, 28, 30, 31, 33, 34, 15], which performs approximate similarity search on multimedia database with significant speedup and acceptable accuracy. See [27] for a comprehensive survey.

Prior cross-modal hashing methods can be categorized into unsupervised methods and supervised methods. IMH [20] and CVH [12] are unsupervised methods, which extend spectral hashing [29] to multimodal scenarios and learn the hash functions using eigenvalue decomposition. CMSSH [3] and SCM [31] are supervised methods, which extend supervised learning methods to fit pairwise labels indicating whether two points are known to be similar or dissimilar (in sense of semantic similarity), and if two points are similar then their corresponding hash codes should also be similar. Since supervised methods can explore the semantic labels for enhancing the cross-modal correlations and reducing the semantic gap [19], they can achieve superior accuracy than unsupervised methods for cross-modal similarity search.

A limitation of previous supervised cross-modal hashing methods is that they cannot exploit *nonlinear* correlation across different modalities. As cross-modal data (e.g. image and text) are heterogeneous in nature, it is unlikely that the correlation across modalities can be captured by shallow models. Furthermore, the latest cross-modal retrieval models via deep learning [21, 28, 6, 25] have shown that deep models can distill nonlinear cross-modal correlations more effectively. Despite the success of deep models in cross-modal search, it remains unclear how to explore both feature correlation and semantic correlation in a deep model for cross-modal hash learning. This work will formally study this problem. While deep convolutional neural networks (CNNs) are powerful for content-based image retrieval [24], how to jointly handle cross-modal retrieval for image and text in CNNs is still an open problem. Hence we will focus on deep autoencoders [23] in this work.

## 3. CORRELATION AUTOENCODER HASHING

### 3.1 Problem Definition

For ease of presentation, we describe the CAH approach with only two modalities (e.g. image and text), which can be readily extended to multiple modalities. In cross-modal search system, the database consists of objects from one modality while the query consists of objects from a different modality. To distill the correlation structure across different modalities, we are given a training set of $n$ bimodal objects $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$ denote the two modalities of feature dimensions $d_x$ and $d_y$, respectively. In supervised cross-modal search, we are further given semantic labels of training objects, that is, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), \boldsymbol{l}_i\}_{i=1}^n$, where $\boldsymbol{l} \in \mathbb{R}^c$ denotes the label vector of a bimodal object, and $c$ is the number of categories. Assume each object is associated with at least one of the $c$ categories, hence $l_{ik} = 1$ if object $i$ is associated with category $k$, and $l_{ik} = 0$ otherwise.

The problem of CAH can be formally defined as jointly learning two hashing functions for the two modalities, i.e. $h_x(\boldsymbol{x}) : \mathbb{R}^{d_x} \mapsto \{-1, 1\}^b$ and $h_y(\boldsymbol{y}) : \mathbb{R}^{d_y} \mapsto \{-1, 1\}^b$, where $b$ is the length of the binary hash code. These hashing functions map feature vectors

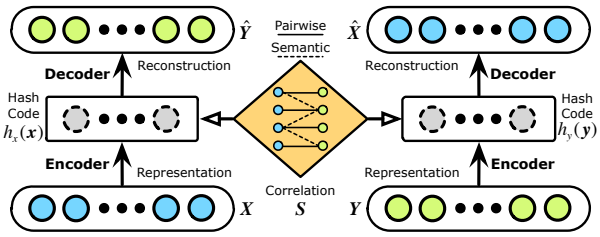



**Figure 1: Correlation Autoencoder Hashing (CAH). Both feature correlation (solid lines) and semantic correlation (dashed lines) are maximized. Note that the features are reconstructed across modalities for feature correlation maximization.**

in the corresponding modality into the common Hamming space, where the similarity between different modalities can be efficiently computed. We adopt the widely-used linear-sign hashing functions defined as $h_x(\boldsymbol{x}) = \operatorname{sgn}\left(\boldsymbol{W}_x^{\mathsf{T}}\boldsymbol{x}\right)$ and $h_y(\boldsymbol{y}) = \operatorname{sgn}\left(\boldsymbol{W}_y^{\mathsf{T}}\boldsymbol{y}\right)$, where $\operatorname{sgn}(\cdot)$ denotes the element-wise sign function, while $\boldsymbol{W}_x \in \mathbb{R}^{d_x \times b}$ and $\boldsymbol{W}_y \in \mathbb{R}^{d_y \times b}$ are the projection matrices. In this paper, we propose to learn these modality-specific hashing functions by maximizing cross-modal *deep* correlations conveyed in *both* feature vectors and semantic labels. An intuitive illustration of the proposed CAH model can be found in Figure 1, which constitutes cross-modal reconstructive embedding and cross-modal semantic correlation.

### 3.2 Cross-modal Reconstructive Embedding

In the context of similarity search, our goal is to rank the most relevant database objects that are similar to the query objects in terms of a pre-defined similarity measure, e.g. Hamming distance. Hence, preserving the similarity information conveyed in original feature vectors serves as an important learning criterion for hashing quality. This criterion can be achieved by minimizing the quantization error of transforming original feature vectors to binary hash codes, or equivalently, minimizing the reconstruction error of transforming binary hash codes to original feature vectors [17, 6]. A common practice is to learn a pair of modality-specific transformation matrices $\boldsymbol{V}_x \in \mathbb{R}^{d_x \times b}$ and $\boldsymbol{V}_y \in \mathbb{R}^{d_y \times b}$ such that the reconstruction errors within each modality are minimized,

$$\min_{\boldsymbol{V}_x, \boldsymbol{V}_y} \sum_{i=1}^{n} \left( \|\boldsymbol{x}_i - \boldsymbol{V}_x h_x(\boldsymbol{x}_i)\|_2^2 + \|\boldsymbol{y}_i - \boldsymbol{V}_y h_y(\boldsymbol{y}_i)\|_2^2 \right), \quad (1)$$

in which $h_x(\boldsymbol{x}) = \operatorname{sgn}\left(\boldsymbol{W}_x^{\mathsf{T}}\boldsymbol{x}\right)$ and $h_y(\boldsymbol{y}) = \operatorname{sgn}\left(\boldsymbol{W}_y^{\mathsf{T}}\boldsymbol{y}\right)$ are the hashing functions. We can observe that Equation (1) takes a similar form as autoencoders [2], but with a different sgn activation function that can learn binary hash codes.

While the reconstructive embedding in Equation (1) can preserve the similarity information within each modality, it may fail to distill the correlation structure across different modalities and restrict the cross-modal search performance. To this end, we propose to reconstruct the original feature vectors from its pairwise hash codes in a different modality. The *cross-modal* reconstructive embedding is formulated as

$$\min_{\boldsymbol{V}_x, \boldsymbol{V}_y} L = \sum_{i=1}^{n} \left( \|\boldsymbol{x}_i - \boldsymbol{V}_x h_y(\boldsymbol{y}_i)\|_2^2 + \|\boldsymbol{y}_i - \boldsymbol{V}_y h_x(\boldsymbol{x}_i)\|_2^2 \right), \quad (2)$$

where $L$ is the aggregate error function. The demonstration of the cross-modal reconstructive embedding can be found in Figure 1. The hash function learning of one modality explores the similarity information from another modality and vice-versa. In this way,

we can capture the cross-modal correlations conveyed in the feature vectors using bimodal data. Note that our work learns binary hashing functions whereas the similar previous work [6] learns continuous feature representations.

## 3.3 Cross-modal Semantic Correlation

The aforementioned unsupervised cross-modal reconstructive embedding may still suffer from two limitations: (1) Without exploring semantic information, we only know the *pairwise correlation* conveyed in bimodal data, i.e. the hash codes of the two modalities within each training object should be similar. Such pairwise correlation in feature vectors is insufficient to distill the cross-modal correlations, since we cannot successfully infer whether unpaired modalities may convey correlations, i.e. whether $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ $(i \neq j)$ are correlated and should have similar embeddings. (2) Due to the well-known *semantic gap* [19] issue, high-level semantic description of an object often deviates from low-level feature descriptors, hence returning nearest neighbors according to similarity measures between original feature vectors cannot always guarantee satisfactory search quality. To address the above two limitations, we need to enhance the cross-modal correlation by exploring semantic information. Previous works usually require objects of the same category to have similar embeddings [26, 13, 31, 30, 32]. However, this requirement is indeed too strict for real-world problems as there is large intra-class variance (e.g. with subclasses) such that objects of the same category may have different embeddings, especially for the objects across different modalities.

In this paper, we propose a novel cross-modal semantic correlation approach by taking the following justifications into consideration: (1) object pairs from different categories should be separated and have discriminative embeddings (this is to maximize the inter-category separation margin); (2) object pairs from the same category should have similar embeddings only if they are similar in the original feature spaces (this is to circumvent the large intra-class variance). Denote by $\boldsymbol{S}^b$ and $\boldsymbol{S}^w$ the between-class and within-class similarity matrices respectively, then the two learning criteria for maximizing the cross-modal semantic correlations are

$$\begin{aligned} &\max_{\boldsymbol{W}_x,\boldsymbol{W}_y} \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^b \left\|h_x\left(\boldsymbol{x}_i\right) - h_y\left(\boldsymbol{y}_j\right)\right\|_2^2, \\ &\min_{\boldsymbol{W}_x,\boldsymbol{W}_y} \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^w \left\|h_x\left(\boldsymbol{x}_i\right) - h_y\left(\boldsymbol{y}_j\right)\right\|_2^2. \end{aligned} \tag{3}$$

The above criterion is similar to Linear Discriminant Analysis (LDA) [22], but with two notable distinctions: (1) the pairwise similarity of embedding is defined on cross-modal binary hash codes $h_x(\boldsymbol{x}_i)$ and $h_y(\boldsymbol{y}_j)$, while LDA defines it on intra-modal continuous embeddings; (2) we also require the locality constraint for intra-class object pairs, which can extract more discriminative and fine-grained hash codes by exploring subclass structures that potentially hide underlying the large intra-class variance.

To formally define the locality-aware similarity matrices, we need to first construct a nearest neighbor affinity matrix $\boldsymbol{A}$ to capture cross-modal locality information as follows,

$$A_{ij} = \begin{cases} d\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right), & \text{if } \boldsymbol{l}_i = \boldsymbol{l}_j \wedge \begin{cases} \boldsymbol{x}_i \in \mathcal{N}_k\left(\boldsymbol{x}_j\right) \vee \boldsymbol{x}_j \in \mathcal{N}_k\left(\boldsymbol{x}_i\right) \\ \boldsymbol{y}_i \in \mathcal{N}_k\left(\boldsymbol{y}_j\right) \vee \boldsymbol{y}_j \in \mathcal{N}_k\left(\boldsymbol{y}_i\right) \end{cases} \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{N}_k(\boldsymbol{x})$ represents the $k$-nearest neighbors of $\boldsymbol{x}$ and $d\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right) = e^{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 / 2\sigma_x^2} + e^{-\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2^2 / 2\sigma_y^2}$ is the cross-modal similarity between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$. Heat kernel parameter $\sigma_x^2 = \frac{1}{n}\sum_{ij}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2$ is the expectation of all pairwise distances in $\{\boldsymbol{x}_i\}_{i=1}^n$, while similar definition applies to $\sigma_y^2$. With the locality affinity matrix $\boldsymbol{A}$, which may be sparse using a practical choice of $k = 10$, we can formulate the similarity matrices involving the semantic information as

$$\begin{aligned} S_{ij}^b &= \begin{cases} A_{ij}\left(1/n - 1/n_c\right), & \text{if } \boldsymbol{l}_i = \boldsymbol{l}_j = c \\ A_{ij}/n, & \text{if } \boldsymbol{l}_i \neq \boldsymbol{l}_j, \end{cases} \\ S_{ij}^w &= \begin{cases} A_{ij}/n_c, & \text{if } \boldsymbol{l}_i = \boldsymbol{l}_j = c \\ 0, & \text{if } \boldsymbol{l}_i \neq \boldsymbol{l}_j, \end{cases} \end{aligned} \tag{5}$$

where $n_c$ is the number of objects within the $c$-th category. The above semantic similarity matrices are consistent with LDA, except for introducing the locality affinity matrix $\boldsymbol{A}$. As can be examined, far apart object pairs in the same category have less influence on $\boldsymbol{S}^b$ and $\boldsymbol{S}^w$ with smaller weight $A_{ij}$. Note that we reweight the values for the object pairs in different categories since we want to separate them more from each other if they are similar in original feature space.

For expression conciseness, we rewrite the cross-modal *semantic correlation* criterion into a unified formulation as

$$\min_{\boldsymbol{W}_x,\boldsymbol{W}_y} R = \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\left\|h_x\left(\boldsymbol{x}_i\right) - h_y\left(\boldsymbol{y}_j\right)\right\|_2^2, \tag{6}$$

where $R$ is the penalty of cross-modal semantic correlation, and the unified semantic similarity matrix $\boldsymbol{S} = \boldsymbol{S}^w - \boldsymbol{S}^b$ is

$$S_{ij} = \begin{cases} A_{ij}\left(2/n_c - 1/n\right), & \text{if } \boldsymbol{l}_i = \boldsymbol{l}_j = c \\ -A_{ij}/n, & \text{if } \boldsymbol{l}_i \neq \boldsymbol{l}_j. \end{cases} \tag{7}$$

It is worth noting that, in the proposed CAH approach, the semantic information is not necessary to be semantic labels. The approach also supports other forms of pairwise labels indicating whether two data points are known to be similar or dissimilar, e.g. the relevance feedback in real-world search engines.

## 3.4 Unified Optimization Problem

An original motivation of this work is to jointly explore both *feature correlation* (2) and *semantic correlation* (6) in a unified learning framework for cross-modal hashing. This motivation leads to the Correlation Autoencoder Hashing (CAH) model, with the joint optimization problem formulated as

$$\begin{aligned} &\min_{\boldsymbol{V}_x,\boldsymbol{V}_y,\boldsymbol{W}_x,\boldsymbol{W}_y} O = L + \lambda R \\ &h_x\left(\boldsymbol{x}\right) = \operatorname{sgn}\left(\boldsymbol{W}_x^{\mathrm{T}}\boldsymbol{x}\right), h_y\left(\boldsymbol{y}\right) = \operatorname{sgn}\left(\boldsymbol{W}_y^{\mathrm{T}}\boldsymbol{y}\right), \end{aligned} \tag{8}$$

where $\lambda$ is a penalty parameter for trading off the relative importance of feature correlation and semantic correlation. In summary, CAH is the first cross-modal hashing method that simultaneously distills feature correlation and semantic correlation in a unified optimization framework: (1) CAH enhances feature correlation by cross-modal reconstructive embedding, which reconstructs the original feature vectors from its pairwise hash codes in another different modality; (2) CAH maximizes the inter-category separation margin for learning more discriminative hash codes; and (3) CAH minimizes the intra-category variance by further exploring the cross-modal locality information, which produces more fine-grained hash codes that characterize the subclass structures of each category to facilitate more accurate similarity ranking. Because different modalities are mapped into the common Hamming space by CAH, it is ready to use the model for cross-modal retrieval, e.g. using a text query to search relevant images from the database.

One of the most difficult issues in hash function learning is that hash codes are binary, making the objective neither continuous nor differentiable, and it is NP-hard to directly compute the best binary

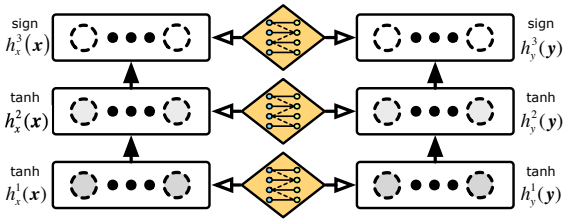



**Figure 2: Deep architecture of the CAH model (only hash code layers are shown). At each layer, CAH maximizes both feature and semantic correlation.**

hash functions of the problem (8). Most of the existing methods hence resort to the continuous relaxation of hash codes for numerical optimization [27]. A widely adopted option is spectral relaxation $h(\boldsymbol{x}) = \boldsymbol{W}^\mathsf{T}\boldsymbol{x}$ [29, 26, 31]. Yet, such relaxation incurs large binarization error as $\boldsymbol{W}^\mathsf{T}\boldsymbol{x} \in (-\infty, +\infty)$ while $\mathrm{sgn}(\boldsymbol{W}^\mathsf{T}\boldsymbol{x}) \in \{-1, 1\}$. In this paper, we are motivated by the success of logistic regression in squashing continues predictions to binary categories of $[-1, 1]$, and adopt the hyperbolic tangent function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ as the reasonable surrogate for $\mathrm{sgn}(x)$. Besides that $\tanh$ and sgn have consistent value ranges of $[-1, 1]$ and guarantee minimized binarization error, another benefit of $\tanh$ is that it injects nonlinearity to the CAH problem (8), making it viable for learning deep hash codes by layerwise abstraction. Detailed learning algorithms are presented in Section 3.6.

## 3.5 Deep Architecture

Cross-modal data have significantly different statistical properties, which makes it very challenging to capture the correlation structures across modalities directly. Recently, it has been witnessed that deep learning methods [2], such as deep autoencoders and convolutional networks, have made performance breakthroughs on many real-world recognition problems. Deep architectures are particularly powerful for distilling the cross-modal correlation since they can extract multilayer feature representations with different abstraction levels [21, 6]. Hence we propose a deep architecture with CAH as building block to enhance cross-modal correlation.

We adopt the stacked autoencoders architecture [28]. After the first layer CAH is trained, we can construct the deep architecture by stacking multiple CAH's on top of it in a layerwise manner, where the *hidden* output of the lower layer is used as the input of the upper layer. We train the Stacked CAH in a greedy layerwise manner by using the Stochastic Gradient Descent (SGD) [23] algorithm, detailed in Section 3.6. An illustration of Stacked CAH is shown in Figure 2. To guarantee accurate hidden representation, we only perform binarization on the top layer to obtain the hash codes. An important advantage of our deep architecture over[6] is that the feature correlation and semantic correlation are distilled in each layer, and can be strengthened layer by layer.

## 3.6 Learning Algorithm

When training each CAH by back-propagation (BP) using mini-batch SGD, we only need to consider the gradient of objective (8) w.r.t. each data point $\boldsymbol{x}_i$ and its correlated points $\boldsymbol{y}_j$'s such that $S_{ij} \neq 0$. CAH can be seen as a neural network with one hidden layer for hash codes. The output layer is similar to standard autoencoders [23] and can be computed in a similar procedure, while the hidden layer is different from standard autoencoders due to the cross-modal semantic correlation regularizer in Equation (6). Hence for each data point $\boldsymbol{x}_i$, we need to compute the gradient of (6) involving $\boldsymbol{x}_i$, i.e. $R(\boldsymbol{x}_i)$ w.r.t. parameters $\boldsymbol{W}_x$. For instance,

$$\begin{aligned}\frac{\partial R(\boldsymbol{x}_i)}{\partial \boldsymbol{W}_{pq}^x} &= \frac{\partial}{\partial \boldsymbol{W}_{pq}^x} \sum_{j=1}^{n} S_{ij} \left\| h(\boldsymbol{W}_x^\mathsf{T}\boldsymbol{x}_i) - h(\boldsymbol{W}_y^\mathsf{T}\boldsymbol{y}_j) \right\|_2^2 \\ &= \sum_{j=1}^{n} 2S_{ij} \left[ h(\boldsymbol{W}_{*q}^{x\mathsf{T}}\boldsymbol{x}_i) - h(\boldsymbol{W}_{*q}^{y\mathsf{T}}\boldsymbol{y}_j) \right] h'(\boldsymbol{W}_{*q}^{x\mathsf{T}}\boldsymbol{x}_i)\boldsymbol{x}_{pi},\end{aligned} \tag{9}$$

where $\boldsymbol{W}_{*q}^x$ is the $q$-th column of $\boldsymbol{W}_x$ and $\boldsymbol{x}_{pi}$ is the $p$-th element of $\boldsymbol{x}_i$, and $h'$ is the gradient of the $\tanh$ function, which is $h'(x) = 1 - (h(x))^2$. Similarly, for each data point $\boldsymbol{y}_i$, the gradient $\frac{\partial R(\boldsymbol{y}_i)}{\partial \boldsymbol{W}_{pq}^y}$ can be computed in the above way.

After getting the gradient of the cross-modal correlation regularizer (6), we can compute the gradient of the overall objective function (8) w.r.t. each point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ as follows,

$$\frac{\partial O(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\partial \boldsymbol{W}_{pq}^x} = \frac{\partial L(\boldsymbol{y}_i)}{\partial \boldsymbol{W}_{pq}^x} + \lambda \frac{\partial R(\boldsymbol{x}_i)}{\partial \boldsymbol{W}_{pq}^x}, \tag{10}$$

where $\frac{\partial L(\boldsymbol{y}_i)}{\partial \boldsymbol{W}_{pq}^x}$ is computed by BP for standard autoencoders.

**Computational Complexity:** The standard autoencoder requires $O(tbdn)$ cost, where $t$ is the number of epochs, $b$ is the number of hidden units, $d$ is the feature dimension, and $n$ is the number of samples. We denote by $z$ the average number of nonzero elements in each row of the semantic similarity matrix $\boldsymbol{S}$, then computing Equation (9) for all $\{\boldsymbol{x}_i\}_{i=1}^n$ requires $O(tb^2dzn)$. The overall computational complexity is $O(tb^2dzn)$, which scales linearly w.r.t. $n$. Though naively building affinity matrix $\boldsymbol{A}$ requires $O(zn^2)$, the cost can be reduced to $O(n \log n)$ by heap structure [10].

**Table 1: The Statistics of the Three Benchmark Datasets**

| Dataset | NUS-WIDE | Wiki | Flickr |
|---|---|---|---|
| Complete Set | 195,834 | 2,866 | 1,000,000 |
| Database | 191,834 | 2,173 | 997,000 |
| Query Set | 2,000 | 393 | 2,000 |
| Training Set | 10,000 | 2,173 | 22,000 |
| Validation Set | 2,000 | 300 | 1,000 |

# 4. EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the proposed CAH model in comparison with several state-of-the-art hashing methods on three public cross-modal datasets. We investigate the widely adopted evaluation criteria including mean average precision (MAP) and precision-recall curve. The code and experimental configurations will be made available online.

## 4.1 Datasets

We conduct our experimental evaluation on three public benchmark datasets, i.e. **NUS-WIDE** [5], **Wiki** [18], and **Flickr** [8]. Detailed statistics are summarized in Table 1.

**NUS-WIDE**[1] is a public Web image dataset containing 269, 648 images downloaded from Flickr, together with the associated raw tags of these images. There are 81 ground truth concepts (categories) manually annotated for search evaluation. Following previous works [35, 28], we prune the original NUS-WIDE dataset to form a new dataset consisting of 195,834 image-text pairs that belong to one of the 21 most frequent concepts. The images are represented by 500-dimensional bag-of-visual words and the texts are represented by 1000-dimensional tag occurrence vectors.

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

**Wiki**[2] is crawled from the Wikipedia's featured articles, with 2,866 image-text pairs. A Wikipedia article contains multiple sections, and the text information and its associated image in one section is considered as an image-text pair, which is annotated by semantic labels of 10 categories. Each image is represented by a 128-dimensional bag-of-words feature vector extracted from SIFT features, and each text is represented by the probability distribution over 10 topics learned by Latent Dirichlet Allocation (LDA).

**Flickr** includes 1 million images associated with tags from Flickr, in which 25,000 are labeled with 38 concepts while the remaining 975,000 are unlabeled. Each image of the public available preprocessed dataset[3] is represented by a 3,857-dimensional vector concatenated by local SIFT feature, color histogram, global GIST feature, and etc [21]. Each text is represented by a 2,000-dimensional feature vector extracted from tag occurrences like NUS-WIDE.

To make results comparable, we adopt the data preprocessing strategy, ZCA whitening [11, 28], to normalize each dimension of the feature to be zero mean and unit variance.

## 4.2 Comparison Methods

We compare CAH with five state-of-the-art cross-modal hashing methods: **IMH**[4] [20], **CVH**[5] [12], **CMSSH**[5] [3], **SCM** [31] and **CorrAE**[6] [6], which can be organized into three categories:

- **Unsupervised shallow hashing: IMH** embeds intra-media and inter-media similarities into a common Hamming space, and integrates a linear regression to learn hash functions such that hash codes for new data can be efficiently generated. **CVH** extends spectral hashing [29] to cross-modal scenario by mapping similar objects across different modalities to similar binary codes.
- **Supervised shallow hashing: CMSSH** is a among the first works which embeds cross-modal data into a common Hamming space via supervised similarity learning. **SCM** is the most recent work which seamlessly integrates semantic labels into the hash function learning procedure for the large-scale data modeling.
- **Unsupervised deep hashing: CorrAE** learns a cross-modal deep autoencoder by integrating unsupervised representation learning and correlation modeling together. It is worth noting that CorrAE is an embedding approach (instead of hashing approach) to cross-modal retrieval, so we directly apply the *sign* function to the CorrAE embedding for generating binary hash codes.

In order to study the effectiveness of our approach with respect to either feature correlation, semantic correlation or data locality, we further evaluate two variants of **CAH** : 1) CAH with only feature correlation, that is $\boldsymbol{S} = \boldsymbol{I}$ in (8), termed **CAH-F**; 2) CAH without using data locality, that is $\boldsymbol{A} = \boldsymbol{1}$ in (8), termed **CAH-L**. All variants use a 3-layer deep architecture.

## 4.3 Evaluation Protocols

We adopt *Mean Average Precision* (MAP) to measure cross-modal search quality, as it is widely adopted in the literature [20, 35, 28, 30, 6]. Given a query and a set of $R$ retrieved documents,

[2] http://www.svcl.ucsd.edu/projects/crossmodal

[3] http://www.cs.toronto.edu/~nitish/multimodal

[4] http://staff.itee.uq.edu.au/shenht/UQ_IMH

[5] http://www.cse.ust.hk/~dyyeung/code/mlbe.zip

[6] https://github.com/fangxiangfeng/deepnet

**Table 2: MAP Results on the NUS-WIDE Dataset**

| Task | Method | NUS-WIDE | | | |
|---|---|---|---|---|---|
| | | 8 bits | 16 bits | 32 bits | 64 bits |
| $I \rightarrow T$ | IMH | 0.4345 | 0.4399 | 0.4203 | 0.4115 |
| | CVH | 0.4227 | 0.4287 | 0.4074 | 0.3999 |
| | CMSSH | 0.3950 | 0.4052 | 0.4076 | 0.3516 |
| | SCM | 0.4693 | 0.4648 | 0.4619 | 0.4851 |
| | CorrAE | 0.4398 | 0.4522 | 0.4699 | 0.4944 |
| | CAH-F | 0.4439 | 0.4711 | 0.4922 | 0.5234 |
| | CAH-L | 0.4880 | 0.5050 | 0.5219 | 0.5581 |
| | CAH | **0.4920** | **0.5084** | **0.5407** | **0.5628** |
| $T \rightarrow I$ | IMH | 0.4380 | 0.4582 | 0.4186 | 0.4051 |
| | CVH | 0.4787 | 0.4689 | 0.4522 | 0.4453 |
| | CMSSH | 0.3783 | 0.3499 | 0.3944 | 0.4015 |
| | SCM | 0.4449 | 0.4859 | 0.5105 | 0.5259 |
| | CorrAE | 0.4303 | 0.4501 | 0.4634 | 0.4880 |
| | CAH-F | 0.4433 | 0.4666 | 0.4885 | 0.5157 |
| | CAH-L | 0.4933 | 0.5053 | 0.5205 | 0.5250 |
| | CAH | **0.5019** | **0.5135** | **0.5451** | **0.5800** |

*Average Precision* (AP) is defined as

$$\text{AP@}R = \frac{\sum_{r=1}^{R} P\left(r\right)\delta\left(r\right)}{\sum_{r'=1}^{R}\delta\left(r'\right)}, \tag{11}$$

where $P(r)$ denotes the precision of the top $r$ retrieved results, and $\delta(r) = 1$ if the $r$-th retrieved result is a true neighbor of the query, otherwise $\delta(r) = 0$, and here we follow literatures [16, 30, 28] to adopt AP@$R = 50$. Then MAP is computed as the mean of all the queries' average precision.

Furthermore, we also report a widely adopted metric, *precision-recall* curve [35, 28] that shows the variation of precision in different recall levels for fine-grained analysis.

The CAH approach only involves one model parameter, penalty coefficient $\lambda$, for trading off the relative importance of feature correlation and semantic correlation. Here we can automatically select $\lambda$ using cross-validation via annotation ground truths in the validation sets (see Table 1). In Section 4.6, we further study parameter sensitivity for $\lambda$ to validate that CAH can consistently outperform the state-of-the-arts with a wide range of parameter configurations. We follow literature convention [6, 28] to set the number of units $u_\ell$ in each layer of the deep autoencoders: $u_{\ell-1} = 2u_\ell$ and the number of units in the last layer is set to hash code length $b$.

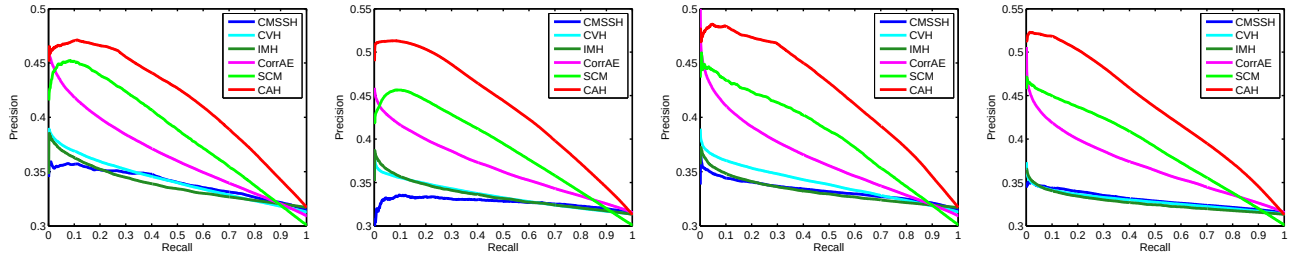
For comparison methods, we also use cross-validation to carefully tune their parameters using the validation sets. Subject to computation burden, it is too costly to train IMH and CMSSH on the complete datasets, hence for fair comparison, we adopt a literature convention [31, 6] and randomly sample 10,000 image-text pairs to train all models. Average results with ten repeated experiments are reported.

## 4.4 Experimental Results

We compare CAH against state-of-the-art methods on cross-modal retrieval tasks (image query on text database $I \rightarrow T$ and text query on image database $T \rightarrow I$) in terms of MAP and precision-recall.

### *4.4.1 Results on NUS-WIDE*

The MAP results of CAH and the comparison methods are demonstrated in Table 2. One can see that CAH achieves significantly better performance on all cross-modal tasks. To zoom-in the CAH model for a deeper understanding on where the performance improvements have come, we further show the MAP results of two




(a) $I \rightarrow T$ @ 16 bits (b) $I \rightarrow T$ @ 32 bits (c) $T \rightarrow I$ @ 16 bits (d) $T \rightarrow I$ @ 32 bits

**Figure 3: Precision-recall curves on the NUS-WIDE cross-modal search tasks $I \rightarrow T$ and $T \rightarrow I$ with hash codes @ 16 and 32 bits.**

**Table 3: MAP Results on the Wiki Dataset**

| Task | Method | Wiki | | | |
|---|---|---|---|---|---|
| | | 8 bits | 16 bits | 32 bits | 64 bits |
| $I \rightarrow T$ | IMH | 0.1734 | 0.1896 | 0.1714 | 0.1601 |
| | CVH | 0.1673 | 0.1877 | 0.1716 | 0.1696 |
| | CMSSH | 0.1672 | 0.1727 | 0.1750 | 0.1759 |
| | SCM | 0.2258 | 0.2372 | 0.2381 | 0.2378 |
| | CorrAE | 0.1990 | 0.2078 | 0.2105 | 0.2177 |
| | CAH-F | 0.2276 | 0.2323 | 0.2233 | 0.2339 |
| | CAH-L | 0.2208 | 0.2342 | 0.2420 | 0.2456 |
| | CAH | **0.2308** | **0.2415** | **0.2465** | **0.2530** |
| $T \rightarrow I$ | IMH | 0.2394 | 0.2227 | 0.2333 | 0.1896 |
| | CVH | 0.2309 | 0.2219 | 0.2214 | 0.2350 |
| | CMSSH | 0.2926 | 0.2991 | 0.2537 | 0.2582 |
| | SCM | 0.3157 | 0.3698 | 0.4239 | 0.4369 |
| | CorrAE | 0.2712 | 0.2948 | 0.3111 | 0.3220 |
| | CAH-F | 0.2608 | 0.3311 | 0.3418 | 0.3693 |
| | CAH-L | 0.3302 | 0.3744 | 0.4156 | 0.4325 |
| | CAH | **0.3424** | **0.3956** | **0.4284** | **0.4569** |

**Table 4: MAP Results on the Flickr Dataset**

| Task | Method | Flickr | | | |
|---|---|---|---|---|---|
| | | 8 bits | 16 bits | 32 bits | 64 bits |
| $I \rightarrow T$ | IMH | 0.5449 | 0.5646 | 0.5936 | 0.5539 |
| | CVH | 0.6091 | 0.6225 | 0.6364 | 0.6199 |
| | CMSSH | 0.5076 | 0.5272 | 0.5357 | 0.5219 |
| | SCM | 0.6361 | 0.6493 | 0.6495 | 0.6440 |
| | CorrAE | 0.6301 | 0.6329 | 0.6357 | 0.6401 |
| | CAH-F | 0.6493 | 0.6470 | 0.6544 | 0.6786 |
| | CAH-L | 0.6520 | 0.6584 | 0.6710 | 0.6920 |
| | CAH | **0.6608** | **0.6875** | **0.7035** | **0.7072** |
| $T \rightarrow I$ | IMH | 0.5374 | 0.5536 | 0.5513 | 0.5583 |
| | CVH | 0.5972 | 0.6032 | 0.5738 | 0.5794 |
| | CMSSH | 0.5868 | 0.5732 | 0.6176 | 0.6323 |
| | SCM | 0.6037 | 0.5998 | 0.5805 | 0.6078 |
| | CorrAE | 0.6142 | 0.6198 | 0.6247 | 0.6431 |
| | CAH-F | 0.6324 | 0.6406 | 0.6508 | 0.6765 |
| | CAH-L | 0.6328 | **0.6734** | **0.6978** | 0.7201 |
| | CAH | **0.6496** | 0.6612 | 0.6908 | **0.7263** |

invariants of CAH: (1) CAH-F achieves worse performance than CAH-L and CAH, which highlights that it is crucial to simultaneously distill feature correlation and semantic correlation for cross-modal search; (2) CAH-L achieves worse performance than CAH, this confirms that it makes no sense to make the far-apart objects in the same category have similar hash codes. By exploring locality-aware semantic similarity, CAH can successfully address the large intra-class issue, which may severely deteriorate search performance but remains a rarely touched issue by previous methods.

An interesting observation is that the unsupervised deep model CorrAE generally performs much better than unsupervised shallow models IMH and CVH. This confirms that cross-modal hashing based on deep models can extract the complex cross-modal correlation more effectively than shallow models. Furthermore, CorrAE even significantly outperforms the conventional supervised method CMSSH and approaches the latest state-of-the-art supervised method SCM. This strongly convinces us the powerfulness of deep models and explains why CAH further outperforms SCM.

The precision-recall curves [35, 28] of all methods are illustrated in Figure 3. CAH establishes the best cross-modal retrieval performance at all recall levels. This validates that CAH is capable for diverse retrieval scenarios, for example, to target a higher recall by tolerating fairly lower precision.

### 4.4.2 *Results on Wiki*

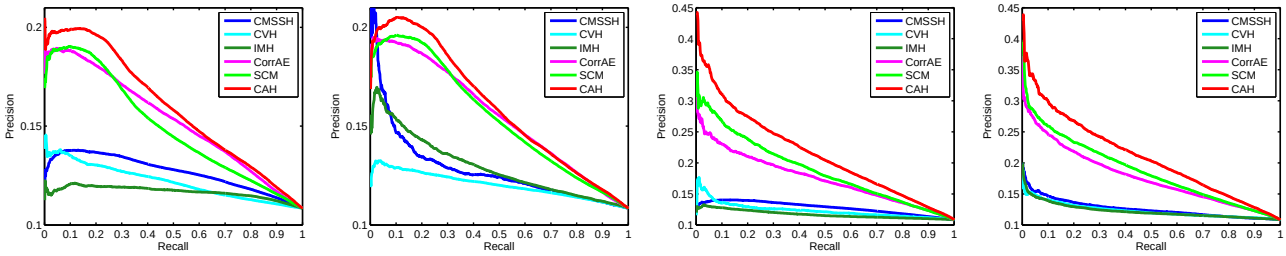
Table 3 shows the MAP scores of CAH and the state-of-arts on the Wiki dataset, which demonstrates that CAH achieves significantly better cross-modal search performance. We also observe that the MAP on $I \rightarrow T$ is substantially lower than the MAP on $T \rightarrow I$. This is an excellent example of the semantic gap issue [19]: the images of Wiki dataset are low-quality and are poorly related to the semantic labels; on the contrary, the texts of Wiki dataset are well-edited featured articles and are more relevant to the semantic labels. We can observe that CAH still performs more effective cross-modal search in the presence of very large semantic gap. The precision-recall curves in Figure 4 show that CAH gives the best cross-modal search quality at all recall levels.

### 4.4.3 *Results on Flickr*

We report the MAP results on the Flickr dataset in Table 4 and show the detailed precision-recall curves in Figure 5. We may observe that CAH can significantly outperform the comparison methods on the two cross-modal search tasks.

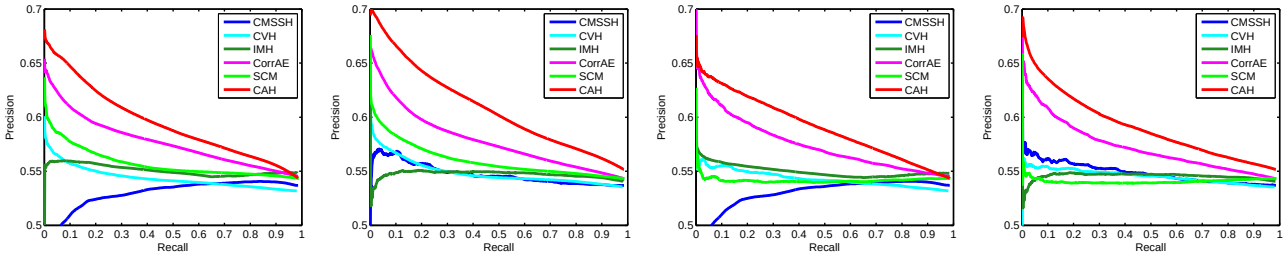
## 4.5 Quantization Error Analysis

The search quality with binary codes in Hamming distance is evidently inferior to searching with continuous features in Euclidean distance, due to substantial information loss by quantizing continuous features to binary codes [28]. Hence, how to minimize the quantization error has been the main effort in the hashing research community. In this regard, we further evaluate IMH, CorrAE and CAH using MAP @ 64 bits with continuous features and binary codes, respectively, with results shown in Figure 6. We can see that CorrAE incurs significantly less loss on search quality than



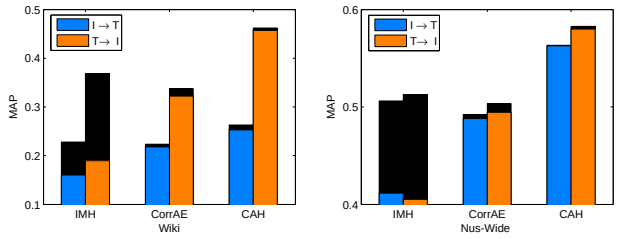

(a) $I \rightarrow T$ @ 16 bits (b) $I \rightarrow T$ @ 32 bits (c) $T \rightarrow I$ @ 16 bits (d) $T \rightarrow I$ @ 32 bits

**Figure 4: Precision-recall curves on the Wiki cross-modal search tasks $I \rightarrow T$ and $T \rightarrow I$ with hash codes @ 16 and 32 bits.**



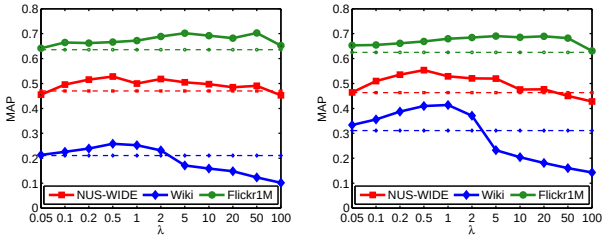

(a) $I \rightarrow T$ @ 16 bits (b) $I \rightarrow T$ @ 32 bits (c) $T \rightarrow I$ @ 16 bits (d) $T \rightarrow I$ @ 32 bits

**Figure 5: Precision-recall curves on the Flickr cross-modal search tasks $I \rightarrow T$ and $T \rightarrow I$ with hash codes @ 16 and 32 bits.**




(a) Quantization Error on Wiki (b) Quantization Error on NUS-WIDE

**Figure 6: Quantization error: search quality loss due to binarization from continuous features to binary codes (black bars).**




(a) $I \rightarrow T$ @ 32 bits (b) $T \rightarrow I$ @ 32 bits

**Figure 7: The MAP @ 32 bits v.s. different values of $\lambda$ for both cross-modal retrieval tasks on NUS-WIDE, Wiki, and Flickr.**

IMH by using binary codes instead of continuous features, while CAH performs even better than CorrAE. This reveals the vital importance of designing good continuous relaxation method for optimization. Our results suggest that the widely-adopted spectral relaxation $\text{sgn}(x) \approx x$ may be too lossy while $\text{sgn}(x) \approx \tanh(x)$ is a very accurate surrogate.

## 4.6 Parameter Sensitivity

The stability of performance against parameter variation is crucial as model selection is time-consuming for large-scale problems. We show in Figure 7 the performance of CAH using MAP @ 32 bits on both cross-modal retrieval tasks by varying $\lambda \in [0.05, 100]$. We see that CAH consistently outperforms the strongest baseline CorrAE on all datasets when $\lambda$ is varied in a large range $[0.1, 2]$.

# 5. CONCLUSION AND FUTURE WORK

In this paper, we have formally approached the problem of supervised cross-modal hashing. The proposed Correlation Autoencoder Hashing (CAH) model simultaneously maximizes the feature correlation revealed by bimodal data and the semantic correlation conveyed in similarity labels, while embeds these correlations into binary hash codes by deep autoencoders. CAH can successfully consolidate the correlation structure across modalities and enhance the generalizability of the embedded hash codes for cross-modal retrieval. Extensive results show that CAH significantly outperforms state-of-the-art cross-modal hashing methods.

In the future, we plan to extend our approach to a hybrid deep learning architecture, which will model text by autoencoders but model image by convolutional neural networks, since each has successful witness in its respective domains. How to seamlessly integrate the heterogeneous deep models for cross-modal hashing remains a very interesting problem.

# Acknowledgments

This work was supported by National Natural Science Foundation of China (61325008, 61502265), China Postdoctoral Science Foundation (2015T80088), National Science and Technology Supporting Program (2015BAH14F02), and Tsinghua National Laboratory (TNList) Special Fund for Big Data Science and Technology.

## 6. REFERENCES


[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *IEEE Symposium on Foundations of Computer Science*, 2006.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, Aug 2013.

[3] M. Bronstein, A. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Computer Vision and Pattern Recognition, 2010 IEEE Conference on*, CVPR'10, pages 3594–3601. IEEE, June 2010.

[4] Y. Cao, M. Long, J. Wang, H. Zhu, and Q. Wen. Deep quantization network for efficient image retrieval. In *AAAI Conference on Artificial Intelligence*, 2016.

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece, July 8-10, 2009.

[6] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.

[7] Y. Hu, Z. Jin, H. Ren, D. Cai, and X. He. Iterative multi-view hashing for cross media indexing. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 527–536. ACM, 2014.

[8] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*. ACM, 2008.

[9] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, Jan 2011.

[10] Z. Jiang, G. Zhang, and L. S. Davis. Submodular dictionary learning for sparse coding. In *CVPR*. IEEE, 2012.

[11] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[12] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI '11, pages 1360–1365, 2011.

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

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

[15] M. Long, Y. Cao, J. Wang, and P. S. Yu. Composite correlation quantization for efficient multimodal retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16. ACM, 2016.

[16] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang. A low rank structural large margin method for cross-modal ranking. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13. ACM, 2013.

[17] M. Norouzi and D. J. Fleet. Cartesian k-means. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13. IEEE, 2013.

[18] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[19] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.

[20] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 785–796. ACM, 2013.

[21] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

[22] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, 8:1027–1061, 2007.

[23] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

[24] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *MM*. ACM, 2014.

[25] D. Wang, P. Cui, M. Ou, and W. Zhu. Deep multimodal hashing with orthogonal regularization. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2015.

[26] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for large-scale search. *TPAMI*, 34(12):2393–2406, 2012.

[27] J. Wang, H. T. Shen, J. Song, and J. Ji. Hashing for similarity search: A survey. Arxiv, 2014.

[28] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. In *Proceedings of the VLDB Endowment*, VLDB '14, pages 649–660. ACM, 2014.

[29] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Advances in Neural Information Processing Systems 21*, pages 1753–1760. Curran Associates, Inc., 2009.

[30] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2014.

[31] D. Zhang and W.-J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[32] P. Zhang, W. Zhang, W.-J. Li, and M. Guo. Supervised hashing with latent factor models. In *SIGIR*, 2014.

[33] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In *Advances in Neural Information Processing Systems 24*, 2012.

[34] Y. Zhen and D.-Y. Yeung. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12. ACM, 2012.

[35] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 143–152. ACM, 2013.