Composite Correlation Quantization for Efficient Multimodal Retrieval*

Mingsheng Long[†], Yue Cao[†], Jianmin Wang[†], and Philip S. Yu[‡♯]
[†]School of Software, Tsinghua National Laboratory (TNList), Tsinghua University, Beijing, China
[‡]Institute for Data Science, Tsinghua University [♯]University of Illinois at Chicago, IL, USA
{mingsheng, jimwang}@tsinghua.edu.cn, caoyue10@gmail.com, psyu@uic.edu

ABSTRACT

Efficient similarity retrieval from large-scale multimodal database is pervasive in modern search engines and social networks. To support queries across content modalities, the system should enable cross-modal correlation and computation-efficient indexing. While hashing methods have shown great potential in achieving this goal, current attempts generally fail to learn isomorphic hash codes in a seamless scheme, that is, they embed multiple modalities in a continuous isomorphic space and separately threshold embeddings into binary codes, which incurs substantial loss of retrieval accuracy. In this paper, we approach seamless multimodal hashing by proposing a novel Composite Correlation Quantization (CCQ) model. Specifically, CCQ jointly finds correlation-maximal mappings that transform different modalities into isomorphic latent space, and learns composite quantizers that convert the isomorphic latent features into compact binary codes. An optimization framework is devised to preserve both intra-modal similarity and inter-modal correlation through minimizing both reconstruction and quantization errors, which can be trained from both paired and partially paired data in linear time. A comprehensive set of experiments clearly show the superior effectiveness and efficiency of CCQ against the state of the art hashing methods for both unimodal and cross-modal retrieval.

Keywords

Hashing, quantization, multimodal retrieval, correlation analysis

1. INTRODUCTION

While big data with large volume, high dimensions, and multiple modalities are ubiquitous in search engines and social networks, it has attracted increasing attention to distill the correlation structures across heterogenous data modalities. For example, an uploaded image on Flickr is usually 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 *multimodal search*, which retrieves semantically-relevant results of all modals in response to a unimodal query. Taking Flickr as an example, when a query image is given, the system should return both relevant tags and images. Due to large volume and semantic gap [18], effective and efficient retrieval of multimodal data remains a challenge.

In the case that 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 based methods [22], which perform approximate nearest neighbor (ANN) search with both computation efficiency and acceptable accuracy. The principle of hashing is to transform high-dimensional data into compact binary codes and generate similar binary codes for similar data items. The seminal work includes Locality Sensitive Hashing (LSH) [1] and Spectral Hashing (SH) [25]. However, traditional *unimodal* hashing methods cannot support multimodal search as ANN cannot be directly computed across different modalities.

Recently, several useful attempts have been made to *multimodal hashing*, which builds correlation structures across multiple modalities in the process of hash function learning and index multimodal data in a common Hamming space [5, 29, 13, 32, 33, 34, 20, 24, 27, 28, 8, 26, 16]. These methods generally work in two-step pipeline: first, embed multiple data modalities into a *continuous* isomorphic latent space by maximizing inter-modal correlations, and second, quantize the isomorphic embeddings into binary hash codes by sign thresholding. While showing promising performance, the two-step pipeline may encounter two limitations: first, conversion from real-valued features to discrete codes may incur substantial information loss, making the continuous latent space suboptimal for binary coding and the binary codes suboptimal for retrieval [24, 10]; second, directly binarizing latent features may lead to unbalanced encoding schemes [32, 33]. Fundamentally, by continuous relaxation of the binary constraints, most methods solve an optimization problem which may deviate significantly from the hashing objective as the quantization error is not accounted for in the optimization process. This somewhat contradicts the motivation of multimodal hashing. Hence, how to learn isomorphic hash codes for multimodal data in a seamless optimization framework remains an open problem.

In this paper, we propose Composite Correlation Quantization (CCQ), a novel model towards seamless multimodal hashing. Technically, CCQ jointly finds correlation-maximal mappings that transform different modalities into an isomorphic latent space, and learns composite quantizers that convert the isomorphic latent features into compact binary codes. The flowcharts of CCQ and prior work are shown in Figure 1. To create a seamless optimization framework, we are inspired by Latent Semantic Analysis (LSA) [7] and decompose each datum into three latent factors, namely, correlation-maximal mapping, similarity-preserving codebook, and compact binary code. The three latent factors are jointly learned through an optimization problem, which preserves both intra-modal similarity

*Corresponding authors: Jianmin Wang and Mingsheng Long.

SIGIR '16, July 17–21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2911493

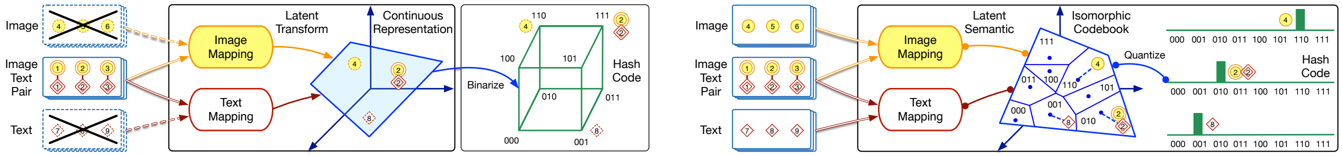


Figure 1: Flowcharts of prior work (left) and CCQ (right). Prior work is a two-step pipeline: first map image-text pairs to isomorphic latent space (denoted as polygon) and then binarize the continuous representation to hash codes (denoted as vertices of hypercube) by *sign thresholding*. CCQ is a seamless optimization framework: jointly map both paired/unpaired images and texts to isomorphic latent space (denoted as polygon) and learn hash codes by *composite quantization*. The quantization model learns isomorphic codebook (denoted as Voronoi digram) and binary codes (denoted as histograms) by minimizing the quantization error, which suffices to assign each latent representation to M-nearest codewords (denoted as Voronoi cells) and assignment indices are used as hash codes.

and inter-modal correlation while minimizing both reconstruction and quantization errors. The CCQ model can construct extremely compressed and balanced binary codes to enable efficient multimodal search, can readily handle a ubiquitous semi-paired scenario where only a fraction of input data are multimodal, and can scale linearly to large sample size. Comprehensive empirical evidence on large-scale datasets confirms that the CCQ model exhibits superior performance in both effectiveness and efficiency on both unimodal and cross-modal search against state of the art hashing methods.

The subsequent paper is organized as follows. We review related works in Section 2. We formally present our model in Section 3 and algorithm with analysis in Section 4. Empirical evaluations are reported in Section 5, while conclusions are enclosed in Section 6.

2. RELATED WORK

Recently, hashing-based multimodal search is a prevalent research focus in machine learning and information retrieval communities [5, 13, 34, 20, 27, 8, 10, 28, 26, 23, 31], which enables approximate similarity search on multimedia database with significant speedup and acceptable accuracy. Refer to [22] for a comprehensive survey.

Existing multimodal hashing methods can be organized into two categories: supervised methods and unsupervised methods. CMSSH [5], SCM [28], QCH [26], and SePH [14] are supervised hashing methods that require labeled pairs to indicate if the objects from different modalities are similar (positive) or dissimilar (negative). As supervised information is usually unavailable in many applications, the deployment of these methods may be severely restricted. CVH [13], IMH [20], MSAE [24] and CorrAE [8] are unsupervised hashing methods applicable to the most general multimodal retrieval case given that paired data are available, while our proposed CCQ model falls into this category. IMH [20] is an extension of spectral hashing [25] to multimodal data, which is restricted by the training burden since constructing and eigendecomposing the similarity matrices require $O(N^2)$. While CVH [13] tackles the scalability issue, it does not jointly maximize cross-modality correlation and preserve intra-modality similarity. MSAE [24] and CorrAE [8] can capture both intra-modal similarity and inter-modal correlation by deep autoencoders, but they require spectral hashing or sign thresholding for obtaining binary codes from the continuous embeddings, which will give rise to uncontrollable quantization errors [9, 12].

A crucial problem with existing methods is that they essentially work in a separated two-step pipeline: first embed multimodal data into a common *continuous* latent space and then threshold the continuous embeddings into binary codes of the Hamming space. Such conversion from real-valued features to discrete codes may result in substantial information loss, making the continuous latent space suboptimal for the binary codes and the binary codes suboptimal for retrieval [30]. Furthermore, directly binarizing latent representation may lead to unbalanced encoding schemes, as shown in [32, 33]. Although IMVH [10] learns multimodal hash functions using a graph-cut quantizer instead of the sign thresholding, the quantizer solves a fast approximation of energy function with orthogonal constraints and recurs large quantization error and unbalanced codes. CCQ approaches this problem by learning the modality-consistent latent space and balanced binary codes in a principled framework.

3. COMPOSITE CORRELATION QUANTIZATION

3.1 Problem Statements

In the multimodal search system, the database and query consist of objects from different modalities. We only use image and text as two modalities to explain our approach, but the approach is formulated to support any number V of modalities. Let $\mathbf{X}^1 \in \mathbb{R}^{P_1 \times N_1}$ be an image set of N_0 images with tags and the rest $\bar{N}_1$ images without tags, where $N_1 = N_0 + \bar{N}_1$ and each image is represented by P_1-dimensional feature vector. Let $\mathbf{X}^2 \in \mathbb{R}^{P_2 \times N_2}$ be a text set of N_0 documents of the image tags and additional $\bar{N}_2$ documents, where $N_2 = N_0 + \bar{N}_2$ and each text is represented by P_2-dimensional feature vector. Note that the proposed approach can handle *semi-paired* data where only a fraction $N_0/(N_1 + N_2)$ of objects are multimodal, and is more realistic than typical multimodal methods.

An efficient approach to calculating the distance between image and text is to map images and texts to modality-isomorphic binary codes in which different modalities of the objects are comparable. In this paper, we will approach this problem by a joint optimization framework, dubbed Composite Correlation Quantization (CCQ).

DEFINITION 1 (CCQ). *Given an image $\mathbf{x}_n^1 \in \mathbb{R}^{P_1}$ and a text $\mathbf{x}_n^2 \in \mathbb{R}^{P_2}$, learn two correlation-maximal mappings $f^1 : \mathbb{R}^{P_1} \mapsto \mathbb{R}^D$ and $f^2 : \mathbb{R}^{P_2} \mapsto \mathbb{R}^D$ that transform images and texts into a D-dimensional isomorphic latent space, and jointly learn two composite quantizers $q^1 : \mathbb{R}^D \mapsto \{0,1\}^H$ and $q^2 : \mathbb{R}^D \mapsto \{0,1\}^H$ that quantize latent embeddings into compact H-bits binary codes.*

In the common H-bits binary space, image and text can be easily comparable such that both intra-modal and cross-modal search can be readily supported. After mappings f^1, f^2 and quantizers q^1, q^2 have been learned, the multimodal search problem can be converted into classical approximate nearest neighbor (ANN) search problem.

3.2 Composite Correlation Quantization

The main idea of CCQ is to jointly learn a correlation-maximal latent space and a similarity-preserving composite quantization in a unified optimization framework. To achieve this mission, we are inspired by Latent Semantic Analysis (LSA) [7] and decompose each input datum (image or text) $\mathbf{x}_n^v$ into three latent factors $\mathbf{R}^v$, $\mathbf{C}^v$, $\mathbf{b}_n^v$, that is, $\mathbf{x}_n^v \approx \mathbf{R}^v \mathbf{C}^v \mathbf{b}_n^v$. While sharing similar formation as LSA, our formulation endows these latent factors with different semantics and thus constrains them with different conditions. More specifically, $\mathbf{R}^v$ is correlation-maximal mapping, $\mathbf{C}^v$ is similarity-preserving codebook, and $\mathbf{b}_n^v$ is the compact binary code of $\mathbf{x}_n^v$. We present how to formulate the CCQ approach under these semantics.

3.2.1 *Intra-Modality Similarity Quantization*

To represent inputs with compact binary codes, two mainstream paradigms are sign thresholding in Hamming embedding methods [25], and vector quantization in codebook-based encoding methods [12]. As sign thresholding cannot guarantee minimal quantization error, we therefore adopt the vector quantization paradigm. CCQ is based on a set of M codebooks $\mathbf{C}^v = [\mathbf{C}_1^v, \ldots, \mathbf{C}_M^v]$, where each codebook $\mathbf{C}_m^v$ contains K codewords $\mathbf{C}_m^v = [\mathbf{C}_{m1}^v, \ldots, \mathbf{C}_{mK}^v]$, and each codeword $\mathbf{C}_{mk}^v$ is a D-dimensional vector like the cluster centroid in kmeans clustering. Corresponding to the M codebooks, we partition the binary codewords assignment vector $\mathbf{b}_n^v$ into M 1-of-K indicator vectors $\mathbf{b}_n^v = [\mathbf{b}_{1n}^v; \ldots; \mathbf{b}_{mn}^v]$, and each indicator vector $\mathbf{b}_{mn}^v$ indicates which one (and only one) of the K codewords in the mth codebook is selected to approximate the nth data point. The CCQ model encodes each $\mathbf{x}_n^v$ as the sum of M codewords, one codeword per codebook, each indicated by the binary assignment vector $\mathbf{b}_n^v$. This yields a novel and more accurate composite approximation scheme $\mathbf{x}_n^v \approx \mathbf{R}^v \sum_{m=1}^{M} \mathbf{C}_m^v \mathbf{b}_{mn}^v$. Consistent with LSA and kmeans, the sum of squared loss between all $\mathbf{x}_n^v$'s and the sum of selected codewords after transformed by $\mathbf{R}^v$, is minimized,

$$\begin{aligned} \min_{\mathbf{R}^v,\mathbf{C}^v,\mathbf{B}^v} \; & \sum_{n=1}^{N_v} \left\| \mathbf{x}_n^v - \mathbf{R}^v \sum_{m=1}^{M} \mathbf{C}_m^v \mathbf{b}_{mn}^v \right\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{b}_{mn}\|_0 = 1, \mathbf{b}_{mn} \in \{0,1\}^K \\ & m = 1 \ldots M, n = 1 \ldots N_v, \end{aligned} \tag{1}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm that simply counts the number of the vector's nonzero elements. The constraint guarantees that only one codeword in each codebook can be activated to approximate the input data, hence it can lead to compact binary codes. As the binary constraints are directly imposed to the learning objective and are valid throughout the optimization procedure, the derived binary codes are much more accurate than sign thresholding binary codes. The rationale of using M codebooks instead of single codebook to approximate each input datum is to further minimize quantization error, as the latter is shown to yield significantly lossy compression and incur evident performance drop [30, 3]. Quantization based on multiple codebooks yields balanced composite binary codes which are more effective than Hamming embedding binary codes [12, 17].

3.2.2 *Inter-Modality Correlation Maximization*

The most desirable value of multimodal retrieval is to enable transfer of knowledge across different modalities so that cross-modal retrieval performance can be improved. A fundamental assumption for multimodal retrieval is that by mapping objects in a modality-consistent latent space, the latent space representations of semantically relevant inter-modal pairs should be consistent. More specifically, for each input object with both image modality $\mathbf{x}_n^1$ and text modality $\mathbf{x}_n^2$, after being transformed by $\mathbf{R}^1$ and $\mathbf{R}^2$ in Equation (1), the latent space representations for image modality $\mathbf{C}^1\mathbf{b}_n^1$ and text modality $\mathbf{C}^2\mathbf{b}_n^2$ should be similar. To our knowledge, most prior work adopts the coupling strategy to minimize $\left\|\mathbf{C}^1\mathbf{b}_n^1 - \mathbf{C}^2\mathbf{b}_n^2\right\|_2^2$. In this paper, we propose to maximize cross-modal correlation by sharing codebooks $\{\mathbf{C}_m\}_{m=1}^M$ for different modalities and sharing binary codes $\{\mathbf{b}_n\}_{n=1}^{N_0}$ for semantically relevant inter-modal pairs. While for the data points with only one modality, the multimodal sharing strategy does not apply. Hence, the proposed condition that the modality-consistent latent space should satisfy is formulated as

$$\mathbf{C}_m^v = \mathbf{C}_m \;\; \text{and} \;\; \delta\left(\mathbf{b}_{mn}^v\right) = \begin{cases} \mathbf{b}_{mn}, & n = 1 \ldots N_0 \\ \mathbf{b}_{mn}^v, & \text{otherwise,} \end{cases} \tag{2}$$

where $\delta(\cdot)$ distinguishes multimodal objects from unimodal ones. Different from most prior methods [20, 8], our modality-consistent condition requires identical code $\mathbf{b}_n^1 = \mathbf{b}_n^2$, instead of minimized distance $\left\|\mathbf{b}_n^1 - \mathbf{b}_n^2\right\|$, for the semantically relevant inter-modal pairs. There are two advantages of our approach. First, since our learning objective keeps the binary constraint valid throughout optimization procedure, it is very difficult to require minimized distance between two binary codes as their nonzero elements may differ significantly. Note that prior methods simply drop the binary condition and solve a continuous problem, which leads to uncontrollable quantization error with the post-step sign thresholding. Second, integrating the minimized distance condition in the learning objective as existing methods may introduce a trade-off term, or parameter, that is hard to tune since its magnitude is very different from learning loss (1).

3.2.3 *Joint Optimization Framework*

To approach CCQ, which jointly learns a similarity-preserving composite quantization and a correlation-maximal latent space in a unified optimization framework, we jointly require the codebooks $\{\mathbf{C}_m\}_{m=1}^M$ subject to minimizing the quantization error of all modalities as Equation (1), and the mappings $\mathbf{R}^v$ subject to maximizing the correlations between semantically relevant inter-modal pairs as Equation (2) while jointly minimizing the reconstruction error of input data as LSA. This leads to a joint optimization framework as

$$\begin{aligned} \min_{\mathbf{R}^v,\mathbf{C},\mathbf{B}^v} \; & \sum_{v=1}^{V}\sum_{n=1}^{N_v} \lambda_v \left\| \mathbf{x}_n^v - \mathbf{R}^v \sum_{m=1}^{M} \mathbf{C}_m \delta\left(\mathbf{b}_{mn}^v\right) \right\|_2^2 \\ \text{s.t.} \quad & \mathbf{R}^{v\mathsf{T}}\mathbf{R}^v = \mathbf{I}_{D\times D}, \mathbf{R}^v \in \mathbb{R}^{P_v \times D} \\ & \left\|\delta\left(\mathbf{b}_{mn}^v\right)\right\|_0 = 1, \delta\left(\mathbf{b}_{mn}^v\right) \in \{0,1\}^K \\ & \delta\left(\mathbf{b}_{mn}^v\right) = \begin{cases} \mathbf{b}_{mn}, & n = 1 \ldots N_0 \\ \mathbf{b}_{mn}^v, & \text{otherwise} \end{cases} \\ & v = 1 \ldots V, m = 1 \ldots M, n = 1 \ldots N_v, \end{aligned} \tag{3}$$

where λ_v is the weight parameter for each modality, and in bimodal problems with $V = 2$, we can simplify the notations by denoting $\lambda_1 = 1$ and $\lambda_2 = \lambda$, while such notations are used throughout this paper. $\mathbf{R}^v$ is the transformation matrix that maps the inputs of each modality to a D-dimensional modality-consistent latent space. The orthogonal constraints are motivated by LSA, which can turn latent factors $\mathbf{R}^v$ into transformation matrices for efficient out-of-sample quantization. The binary codes $\mathbf{b}_n^v$ are $M \times K$-dimensional, fortunately however, each $\mathbf{b}_{mn}^v$ is 1-of-K encoding with only one nonzero element and can be represented using $\log_2 K$ bits, hence the final hash codes $\mathbf{b}_n^v$ can be compacted into $H = M\log_2 K$ bits, which is independent on the dimensions of input or latent spaces. To fit each $\mathbf{b}_{mn}^v$ into one byte, $K = 256$ is a good choice [12, 30]. We simply set $D = \min(\{P_v\}_{v=1}^V, H)$, in the sense that an H-bit binary code can reconstruct a D-dimensional vector accurately.

Formally, we derive correlation-maximal mappings $f^v\left(\mathbf{x}_n^v\right) = \mathbf{R}^{v\mathsf{T}}\mathbf{x}_n^v$ and similarity-preserving quantizers $q^v\left(f^v\left(\mathbf{x}_n^v\right)\right) = \mathbf{b}_n^v$. There are several advantages of the CCQ approach. First, CCQ jointly learns a correlation-maximal latent space and a similarity-preserving composite encoding, which can minimize the quantization loss and guarantee search quality. Second, CCQ explores both paired and unpaired data in a semi-paired quantization paradigm, which can benefit from semi-supervised learning in that paired data consolidate inter-modality correlation and unpaired data enhance intra-modality quantization. Third, CCQ is formulated with only two easy-tuning model parameters D and λ, where D can be set as simply as LSA to retain most covariance information, and λ can be selected by trading off different modalities using prior information. In particular, the proposed sharing of codebooks and binary codes across modalities (2) enables joint learning of latent semantics that are maximally correlated in the isomorphic feature space, which contributes most significantly to the efficacy of the CCQ approach.

3.3 Approximate Nearest Neighbor Search

Approximate nearest neighbor (ANN) search based on Euclidean distance is a powerful task for quantization techniques [12]. Given a database of CCQ hash codes $\{\mathbf{b}_n^v\}_{n=1}^{N_v}$, we follow [12, 17] and use *Asymmetric Quantizer Distance* (AQD) as similarity metric that computes the distance between query $\mathbf{q}^{\bar{v}}$ and database point $\mathbf{x}_n^v$ as

$$\begin{aligned} \mathrm{AQD}\left(\mathbf{q}^{\bar{v}}, \mathbf{x}_n^v\right) &= \left\|\mathbf{q}^{\bar{v}} - \mathbf{R}^{\bar{v}} \sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mn}^v\right\|_2^2 \\ &= -2 \sum\nolimits_{m=1}^{M} \left\langle \tilde{\mathbf{q}}^{\bar{v}}, \mathbf{C}_m \mathbf{b}_{mn}^v \right\rangle + \left\| \sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mn}^v \right\|_2^2 \\ &+ \left\|\tilde{\mathbf{q}}^{\bar{v}}\right\|_2^2 + \left\|\mathbf{R}_{\perp}^{\bar{v}\mathsf{T}} \mathbf{q}^{\bar{v}}\right\|_2^2, \end{aligned} \tag{4}$$

where $\tilde{\mathbf{q}}^{\bar{v}} = \mathbf{R}^{\bar{v}\mathsf{T}} \mathbf{q}^{\bar{v}}$ is the transformed query. In the second row, the first term computes the inner products between $\tilde{\mathbf{q}}^{\bar{v}}$ and M codewords selected by $\mathbf{b}_n^v$. Given a query, these inner products for all M codebooks $\{\mathbf{C}_m\}_{m=1}^M$ and all K possible values of $\mathbf{b}_{mn}^v$ can be pre-computed and stored in a query-specific $M \times K$ lookup table, which is used to compute AQD between the query and all database points, each entails M table lookups and additions and is slightly more costly than Hamming distance. The second term computes the squared norm of decoded database point, which is independent on the query and can be encoded using one byte by quantizing these scale values on held-out dataset [3]. At quantization, we augment CCQ code with the norm byte, which costs one more lookup and one more byte per database point. We can eliminate this norm byte by composite quantization [30], but will leave it to our future work.

4. ALGORITHM AND ANALYSIS

4.1 Learning Algorithm

The CCQ optimization problem (3) consists of three variables, $\mathbf{R}^v$, $\mathbf{C}$, and $\mathbf{B}^v$. We adopt alternating optimization [12, 17, 3, 30] which iteratively updates one variable with the rest variables fixed.

4.1.1 Update $\mathbf{R}^v$

We update $\mathbf{R}^v$ by fixing $\mathbf{C}$ and $\mathbf{B}^v$ as known variables, and write Equation (3) with $\mathbf{R}^v$ as unknown variables in matrix formulation,

$$\begin{aligned} &\min\nolimits_{\mathbf{R}^v} \left\|\mathbf{X}^v - \mathbf{R}^v \mathbf{C} \delta\left(\mathbf{B}^v\right)\right\|_F^2 \\ &\text{s.t.} \quad \mathbf{R}^{v\mathsf{T}} \mathbf{R}^v = \mathbf{I}_{D \times D}. \end{aligned} \tag{5}$$

This is equivalent to the Orthogonal Procrustes problem [19] and can be solved exactly using SVD. More specifically, we perform SVD as $\mathbf{X}^v [\mathbf{C} \delta\left(\mathbf{B}^v\right)]^{\mathsf{T}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$, then we achieve $\mathbf{R}^v = \mathbf{U}\mathbf{V}^{\mathsf{T}}$.

4.1.2 Update $\mathbf{C}$

We update $\mathbf{C}$ by fixing $\mathbf{R}^v$ and $\mathbf{B}^v$ as known variables, and write Equation (3) with $\mathbf{C}$ as unknown variables in matrix formulation,

$$\min_{\mathbf{C}} \sum\nolimits_{v=1}^{V} \left\|\mathbf{R}^{v\mathsf{T}} \mathbf{X}^v - \mathbf{C} \delta\left(\mathbf{B}^v\right)\right\|_F^2. \tag{6}$$

Algorithm 1: CCQ: Composite Correlation Quantization

```
Input: Data $\{\mathbf{X}^v\}_{v=1}^V$; latent dimension $D$, modal weight $\lambda$.
Output: Mappings $\{\mathbf{R}^v\}$, codebook $\mathbf{C}$, binary codes $\{\mathbf{B}^v\}$.
Initialize $\{\mathbf{R}^v\}$ by identity, $\mathbf{C}$ randomly, $\{\mathbf{B}^v\}$ by NN search.
repeat
    Update $\{\mathbf{R}^v\}$ by Orthogonal Procrustes as Eqn. (5).
    Update $\mathbf{C}$ by Quadratic Optimization as Eqn. (6).
    for $n \leftarrow 1$ to $N_v$ do
        Update $\{\mathbf{b}_n^v\}$ by ICM or greedy algorithm as Eqn. (7).
    end
until Convergence
```

This is an unconstrained quadratic problem with analytic solution $\mathbf{C} = \left[\sum\nolimits_{v=1}^{V} \lambda_v \mathbf{R}^{v\mathsf{T}} \mathbf{X}^v \delta\left(\mathbf{B}^v\right)^{\mathsf{T}}\right] \left[\sum\nolimits_{v=1}^{V} \lambda_v \delta\left(\mathbf{B}^v\right) \delta(\mathbf{B}^v)^{\mathsf{T}}\right]^{-1}$. Algorithms such as L-BFGS can be used to speed up computation.

4.1.3 Update $\mathbf{B}^v$

It is obvious that each $\mathbf{b}_n^v$ is independent on $\{\mathbf{b}_{n'}^v\}_{n' \neq n}$, then the optimization problem for $\mathbf{B}^v$ is decomposed to N_v subproblems,

$$\begin{aligned} &\min_{\mathbf{b}_n^v} \sum\nolimits_{v=1}^{V} \lambda_v \left\|\mathbf{R}^{v\mathsf{T}} \mathbf{x}_n^v - \sum\nolimits_{m=1}^{M} \mathbf{C}_m \delta\left(\mathbf{b}_{mn}^v\right)\right\|_2^2 \\ &\text{s.t.} \quad \left\|\delta\left(\mathbf{b}_{mn}^v\right)\right\|_0 = 1, \delta\left(\mathbf{b}_{mn}^v\right) \in \{0,1\}^K. \end{aligned} \tag{7}$$

This optimization problem is generally NP-hard. As shown in [30], this problem is essentially high-order Markov Random Field (MRF) problem and can be solved by the Iterated Conditional Modes (ICM) algorithm [4] which solves M indicators $\{\mathbf{b}_{mn}^v\}_{m=1}^M$ alternatively. Given $\{\mathbf{b}_{m'n}^v\}_{m' \neq m}$ fixed, we update $\mathbf{b}_{mn}^v$ by exhaustively checking all the codeword in codebook $\mathbf{C}_m$, finding the codeword such that the objective in (7) is minimized, and setting the corresponding entry of $\mathbf{b}_{mn}^v$ as 1 and the rest as 0. The algorithm is guaranteed to converge, and can be terminated if maximum iterations are reached. To accelerate quantization, we can explore hierarchical structure of codebooks $\{\mathbf{C}_m\}$ and update $\{\mathbf{b}_{mn}^v\}$ by a new greedy algorithm. Specifically, after updating $\{\mathbf{b}_{m'n}^v\}_{m'<m}$, we can update $\mathbf{b}_{mn}^v$ by encoding residual $\mathbf{R}^{v\mathsf{T}} \mathbf{x}_n^v - \sum\nolimits_{m'=1}^{m-1} \mathbf{C}_{m'} \delta\left(\mathbf{b}_{m'n}^v\right)$ with codebook $\mathbf{C}_m$. The overall learning procedure is summarized in Algorithm 1.

4.2 Large-Scale Implementation

Batch algorithms are memory-inefficient for large-scale datasets, hence we formulate CCQ optimization into mini-batch algorithms for large-scale problems [24]. The main idea is to split the training set into mini-batches and load a fraction of data points into memory each time. Hence, the memory usage stays constant when the size of the training set increases. The update of $\mathbf{B}^v$ in Equation (7) is already mini-batch in that update of each data point is independent on the other data points. To update $\mathbf{R}^v$ in mini-batch, we notice that the matrix for SVD is $\mathbf{X}^v [\mathbf{C} \delta\left(\mathbf{B}^v\right)]^{\mathsf{T}} \in \mathbb{R}^{P_v \times D}$, which if given, the SVD can be solved in $O(P_v^2 D)$, independent on the number of data points. We thus formulate the matrix for SVD in a point-wise summation form as $\sum\nolimits_{n=1}^{N_v} \mathbf{x}_n^v [\mathbf{C} \delta\left(\mathbf{b}_n\right)]^{\mathsf{T}}$, then it can be computed by traversing all data points in a mini-batch paradigm. Similarly, the update of $\mathbf{C}$ can also be formulated in a summation form for mini-batch implementation. Note that we can allocate all available memory to mini-batch and trade off memory and disk reading costs.

4.3 Computational Complexity

We analyze the cost of each iteration to show CCQ scales linearly to sample size N_v. To update $\mathbf{R}^v$, it takes $O\left(N_v P_v D + N_v D M\right)$ to prepare the problem and $O\left(P_v^2 D + D^3\right)$ to compute the SVD.

To update $\mathbf{C}$, it takes $O\left(N_vP_vD + N_vDM + N_vM^2\right)$ to prepare the problem and $O\left(DM^2K^2 + M^3K^3\right)$ to compute the quadratic optimization. To update $\mathbf{B}^v$, it takes $O\left(N_vP_vD + N_vDMKT_i\right)$, where T_i is the number of iterations and $T_i = 3$ in ICM algorithm or $T_i = 1$ in greedy algorithm can obtain satisfactory performance. As a rule of thumb, $D = H$ and $K = 256$ are good choices for most applications. For longer codes, update of $\mathbf{C}$ is inefficient, in which case we can adopt the online L-BFGS algorithm for speedup.

4.4 Approximation Error Analysis

Given a query $\mathbf{q}^{\bar{v}}$ and a database point $\mathbf{x}_n^v$, after transformed by correlation-maximal mappings $\tilde{\mathbf{q}}^{\bar{v}} = \mathbf{R}^{\bar{v}\mathsf{T}}\mathbf{q}^{\bar{v}}$ and $\tilde{\mathbf{x}}_n^v = \mathbf{R}^{v\mathsf{T}}\mathbf{x}_n^v$, they can be comparable in the modality-consistent latent space, and their Euclidean distance is computed as $d\left(\tilde{\mathbf{q}}^{\bar{v}}, \tilde{\mathbf{x}}_n^v\right) = \left\|\tilde{\mathbf{q}}^{\bar{v}} - \tilde{\mathbf{x}}_n^v\right\|_2$. As computing Euclidean distance on real-valued vectors is too costly for large-scale search, we compute AQD (4) on binary codes. Hence, we need to analyze the error bound of using AQD to approximate real-valued distance. Denote $\hat{\mathbf{x}}_n^v = \sum_{m=1}^M \mathbf{C}_m\mathbf{b}_{mn}^v$ the decoded vector of $\mathbf{x}_n^v$, then AQD $\left(\mathbf{q}^{\bar{v}}, \mathbf{x}_n^v\right) = d\left(\tilde{\mathbf{q}}^{\bar{v}}, \hat{\mathbf{x}}_n^v\right) + \epsilon$, ϵ is a constant.

THEOREM 1 (BOUND). *The error is bounded by learning loss*

$$\left|d\left(\tilde{\mathbf{q}}^{\bar{v}}, \tilde{\mathbf{x}}_n^v\right) - d\left(\tilde{\mathbf{q}}^{\bar{v}}, \hat{\mathbf{x}}_n^v\right)\right| \leqslant \left\|\mathbf{x}_n^v - \mathbf{R}^v \sum_{m=1}^M \mathbf{C}_m\mathbf{b}_{mn}^v\right\|_2. \tag{8}$$

PROOF. From the triangle inequality, $\left|d\left(\tilde{\mathbf{q}}^{\bar{v}}, \tilde{\mathbf{x}}_n^v\right) - d\left(\tilde{\mathbf{q}}^{\bar{v}}, \hat{\mathbf{x}}_n^v\right)\right| \leqslant d\left(\tilde{\mathbf{x}}_n^v, \hat{\mathbf{x}}_n^v\right)$. Then

$$\begin{aligned} d^2\left(\tilde{\mathbf{x}}_n^v, \hat{\mathbf{x}}_n^v\right) &= \left\|\mathbf{R}^{v\mathsf{T}}\mathbf{x}_n^v - \sum_{m=1}^M \mathbf{C}_m\mathbf{b}_{mn}^v\right\|_2^2 \\ &\leqslant \left\|\mathbf{R}^{v\mathsf{T}}\mathbf{x}_n^v - \sum_{m=1}^M \mathbf{C}_m\mathbf{b}_{mn}^v\right\|_2^2 + \left\|\mathbf{R}_\perp^{v\mathsf{T}}\mathbf{x}_n^v\right\|_2^2 \\ &= \left\|\mathbf{x}_n^v - \mathbf{R}^v \sum_{m=1}^M \mathbf{C}_m\mathbf{b}_{mn}^v\right\|_2^2, \end{aligned} \tag{9}$$

where $\mathbf{R}_\perp^v$ is an orthogonal complement of $\mathbf{R}^v$, $\mathbf{R}^{v\mathsf{T}}\mathbf{R}_\perp^v = \mathbf{0}$. □

The theorem confirms that the error of using AQD to approximate real-valued distance is statistically bounded by CCQ learning loss. Hence, CCQ is more accurate than sign thresholding methods [25]. An important advantage of CCQ in Equation (9) is that mapping $\mathbf{R}^v$ is learned by a joint optimization of canonical correlation analysis (CCA) and principal component analysis (PCA) corresponding to the first and second terms of Line 2 in Equation (9). This can be much more effective than most CCA-based methods [13, 28, 26].

5. EXPERIMENTS

We conduct extensive evaluation of CCQ against state of the art methods on three public multimodal datasets. We investigate both effectiveness and efficiency in terms of search precision, recall, and time. The codes, data, and configurations will be available online.

5.1 Datasets

The evaluation is conducted on three datasets: NUS-WIDE [6], Wiki [18], and Flickr1M [11], with statistics depicted in Table 1. We preprocess all datasets by applying ZCA [24] to normalize each dimension of image/text features to be zero mean and unit variance.

NUS-WIDE[1] is a Web image dataset containing 269, 648 images downloaded from Flickr, each associated with 6 tags on average. There are 81 ground truth concepts manually annotated for search evaluation. Following prior works [34, 24], we prune the original NUS-WIDE to form a new dataset consisting of 195,834 image-text pairs by keeping the pairs that belong to one of the 21 most frequent concepts. The images are represented by 500-dimensional bag-of-words vectors extracted from the SIFT features using k-means, and the texts are represented by 1,000-dimensional vectors extracted from the tag occurrence features using PCA. A query set of 2,000 image-text pairs are randomly sampled from the dataset, while the remaining 193,834 image-text pairs are serving as the database. The hash models are learned on the training set containing 10,000 image-text pairs randomly sampled from the database [34, 20].

Table 1: The Statistics of Three Datasets

Dataset	NUS-WIDE	Wiki	Flickr1M
Complete Set	195,834	2,866	1,000,000
Labeled Set	195,834	2,866	25,000
Query Set	2,000	693	1,000
Database	193,834	2,173	24,000
Training Set	10,000	2,173	975,000

Wiki[2] contains 2,866 image-text pairs selected from Wikipedia's featured articles comprised of multiple sections of images and texts. Every image-text pair is labeled by one of the 10 concepts in the article categories. Each image is represented by a 128-dimensional bag-of-words vector extracted from SIFT features, and each text is represented by the probability distribution over 10 topics learned by a latent Dirichlet allocation (LDA) model. The dataset is released with a query set of 693 pairs and a database of 2,173 pairs, and the whole database is used as the training set for hash coding [18, 34].

Flickr1M comprises 1,000,000 images associated with tags from Flickr, in which 25,000 are labeled with 38 concepts while the remaining 975,000 are unlabeled. The public available preprocessed dataset[3] is employed for evaluation, in which each image is represented by a 3,857-dimensional vector concatenated by local SIFT feature, global GIST feature, etc [21]. Each text is represented by a 2,000-dimensional vector extracted from tag occurrences. The query set contains 1,000 image-text pairs randomly sampled from the 25,000 labeled pairs, and the rest 24,000 labeled pairs are used as the database. In scalability test of CCQ (Section 5.7), all 975,000 unlabeled pairs are used as the training set for learning hash codes.

5.2 Comparison Methods

We compare CCQ against many state of the art hashing methods.

- **Unsupervised hashing:** Cross-View Hashing (**CVH**)[6] [13] and Inter-Media Hashing (**IMH**)[4] [20] are unsupervised hashing methods that extend spectral hashing to exploit the local structure of multimodal data for learning binary codes.
- **Deep hashing:** Correspondence Auto-Encoders (**CorrAE**)[5] [8] learns latent features via unsupervised deep auto-encoders, which captures both intra-modal and inter-modal correspondences, and binarizes latent features via sign thresholding.
- **Supervised hashing:** Cross-Modal Similarity-Sensitive Hashing (**CMSSH**)[6] [5], Semantic Correlation Maximization (**SCM**) [28], and Quantized Correlation Hashing (**QCH**) are supervised hashing methods which embed multimodal data into a common Hamming space using supervised metric learning.

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

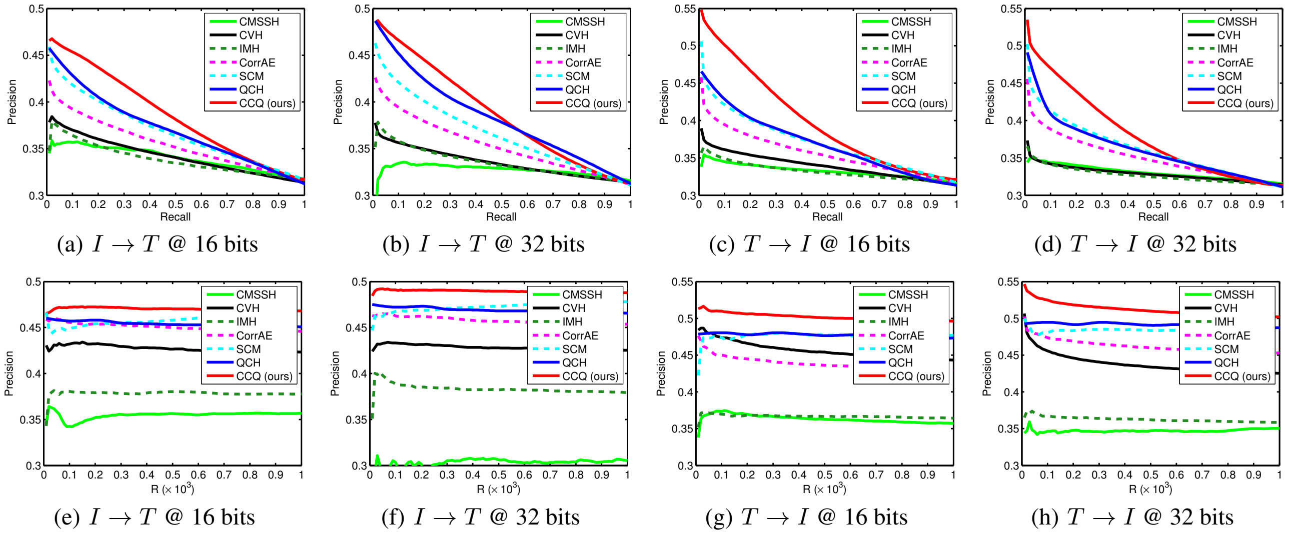
[2] http://www.svcl.ucsd.edu/projects/crossmodal

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

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

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

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



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

(e) $I \to T$ @ 16 bits (f) $I \to T$ @ 32 bits (g) $T \to I$ @ 16 bits (h) $T \to I$ @ 32 bits

Figure 2: Precision-recall curves (top) and precision@R curves (bottom) on NUS-WIDE cross-modal search tasks @ 16 and 32 bits.

5.3 Evaluation Protocols

We perform four types of multimodal retrieval schemes: (1) $I \to I$: use image queries to retrieve relevant images; (2) $T \to T$: use text queries to retrieve relevant texts; (3) $I \to T$: use image queries to retrieve relevant texts; and (4) $T \to I$: use text queries to retrieve relevant images. The first two tasks are intra-modal retrieval and the last two tasks are cross-modal retrieval. As CCQ can also handle multimodal search where both modalities are available for the database, we show the results of multimodal retrieval schemes where each image-text pair is quantized into a unified hash code by fusing knowledge of different modalities: (5) $I \to IT$: use image queries to retrieve relevant image-text pairs; (6) $T \to IT$: use text queries to retrieve relevant image-text pairs. The baseline methods do not support multimodal search because they do not use shared coding for different modalities of the same object. Given a query, the ground truth is defined as: if a result shares at least one common concept with the query, it is relevant; otherwise it is irrelevant.

We adopt *Mean Average Precision* (MAP) to measure the effectiveness of multimodal search [20, 34, 24, 27, 8]. Given a set of queries, we first calculate Average Precision (AP) of each query as

$$\mathrm{AP}@R = \frac{\sum_{r=1}^{R} P(r)\,\delta(r)}{\sum_{r'=1}^{R} \delta(r')}, \tag{10}$$

where R is the number of retrieved documents, $P(r)$ 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 $\delta(r) = 0$. Then MAP is computed as the mean of all the queries' average precision, and the larger the MAP, the better the retrieval performance. In the experiments, we follow [15, 27, 24] to report MAP@$R = 50$. We also report another two standard retrieval criteria, *precision-recall* curves and *precision@top-R* curves of all retrieval tasks. In addition to effectiveness, we report *time and memory* costs as the efficiency measures for query processing and model training.

The CCQ approach involves two model parameters: dimension of modality-consistent subspace D and modality trade-off weight λ. In principle, CCQ is almost immune to different choices of D, as long as D is large enough to retain the majority amount of covariance information as LSA. While no prior knowledge is available, we can simply set equal weights $\lambda = 1$ for different modalities, which can already achieve satisfactory performance. Nonetheless, for image-text bimodal search, the text modality usually carry more semantic information, hence we equip CCQ with the flexibility for selecting the optimal λ to encode such important prior knowledge. Given annotation ground truths as in the evaluation datasets, we can automatically select D and λ using cross-validation. However, we choose to blindly fix $\lambda = 5$ throughout the comparative study. This is desirable as cross-validation may be impossible in the pervasive unsupervised multimodal search. We will study parameter sensitivity in Section 5.8 to validate that CCQ can consistently outperform the state of the arts with a wide range of parameter configurations.

For the comparison methods, we adopt cross-validation to select their optimal parameters, respectively. As cross-validation requires annotation ground truths, this further confirms CCQ's superior parameter stability. Subject to computation burden, it is too costly to train CMSSH and IMH on the complete Flickr1M dataset, hence we randomly sample 10,000 image-text pairs to train these models. Each experiment repeats ten runs and the average result is reported.

5.4 Experimental Results

We compare CCQ with state of the art methods in terms of MAP and precision-recall on 4 multimodal retrieval tasks ($I \to I, T \to T, I \to T, T \to I$) of three datasets (NUS-WIDE, Wiki, and Flickr1M).

5.4.1 Results on NUS-WIDE

We evaluate CCQ against state of the arts with different lengths of hash codes, i.e. 8, 16, 32, and 64 bits, and report the MAP results in Table 2. For all multimodal retrieval tasks, CCQ achieves significantly better performance than all unsupervised hashing methods CVH, IMH, and CorrAE, and generally outperforms the state of the art supervised hashing methods CMSSH, SCM, QCH in most cases. It is very worth noting that, CCQ is an *unsupervised* hashing method that does not require labeled similarity information. Hence CCQ is particularly beneficial when labeled information is unavailable, which is the most common scenario in big data era. A notable limitation of orthogonal constrained methods CVH and IMH is that longer codes do not necessarily improve performance in cross-modal tasks $I \to T$ and $T \to I$. The reason is that these methods learn uncorrelated hash bits via eigenvalue decomposition on similarity matrix, which leads to unbalanced hash codes with the first k eigenvectors (hash bits) dominating the whole hash codes. CCQ

Table 2: Mean Average Precision (MAP) Comparison of Six Multimodal Retrieval Tasks on Three Standard Datasets

Task	Method	NUS-WIDE				Wiki				Flickr1M			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
$I \to I$	CVH [13]	0.3954	0.4542	0.4759	0.4780	0.1988	0.1969	0.2042	0.2058	0.6050	0.6328	0.6615	0.6712
	IMH [20]	0.4313	0.4545	0.4155	0.4005	0.1910	0.1963	0.1937	0.1935	0.5239	0.5725	0.5736	0.5748
	CorrAE [8]	0.4223	0.4478	0.4587	0.4796	0.2055	0.2086	0.2188	0.2194	0.6145	0.6397	0.6588	0.6654
	CMSSH [5]	0.3776	0.4060	0.4356	0.4490	0.1987	0.1979	0.2007	0.2126	0.5738	0.6304	0.6587	0.6932
	SCM [28]	0.4258	0.4578	0.4695	0.4831	0.2048	0.2103	0.2177	0.2212	0.5926	0.6257	0.6615	0.6801
	QCH [26]	0.4289	0.4557	0.4786	0.4898	0.2087	0.2155	0.2198	0.2252	0.6165	0.6586	0.6787	0.6885
	CCQ (ours)	**0.4711**	**0.4859**	**0.4921**	**0.4932**	**0.2226**	**0.2265**	**0.2373**	**0.2386**	**0.6714**	**0.7092**	**0.7318**	**0.7451**
$T \to T$	CVH [13]	0.5825	0.6485	0.6837	**0.7189**	0.4049	0.5506	0.6075	0.6239	0.5812	0.6085	0.6242	0.6337
	IMH [20]	0.4531	0.4740	0.5421	0.6202	0.3805	0.4623	0.5773	0.5989	0.5585	0.5973	0.6360	0.6436
	CorrAE [8]	0.5501	0.5856	0.6344	0.6678	0.5765	0.5889	0.6045	0.6123	0.6060	0.6176	0.6389	0.6443
	CMSSH [5]	**0.5911**	0.5968	0.6215	0.6613	0.5503	0.6065	0.6188	0.6232	0.5487	0.5573	0.5583	0.5614
	SCM [28]	0.5524	0.6315	0.6606	0.6736	0.5814	0.6051	0.6189	0.6324	0.5924	0.6320	0.6410	0.6485
	QCH [26]	0.5706	**0.6586**	0.6796	0.6855	0.6002	0.6128	0.6226	0.6355	0.6022	0.6427	**0.6554**	**0.6686**
	CCQ (ours)	**0.5913**	0.6481	**0.6917**	0.7069	**0.6017**	**0.6286**	**0.6366**	**0.6422**	**0.6090**	**0.6433**	0.6541	0.6550
$I \to T$	CVH [13]	0.4588	0.4713	0.4743	0.4740	0.1673	0.1877	0.1716	0.1696	0.6091	0.6225	0.6364	0.6199
	IMH [20]	0.4345	0.4399	0.4203	0.4115	0.1734	0.1896	0.1714	0.1601	0.5449	0.5646	0.5936	0.5539
	CorrAE [8]	0.4398	0.4522	0.4699	0.4964	0.1929	0.1982	0.2033	0.2155	0.6301	0.6329	0.6357	0.6401
	CMSSH [5]	0.3950	0.4052	0.4076	0.3516	0.1672	0.1727	0.1750	0.1759	0.5076	0.5272	0.5357	0.5219
	SCM [28]	0.4693	0.4648	0.4619	0.4851	0.2258	0.2372	0.2381	0.2378	0.6361	0.6493	0.6495	0.6440
	QCH [26]	0.4765	0.4895	0.5050	0.5125	0.2288	0.2343	0.2368	**0.2402**	0.6452	0.6523	0.6685	0.6721
	CCQ (ours)	**0.5124**	**0.5161**	**0.5165**	**0.5372**	**0.2338**	**0.2349**	**0.2371**	0.2374	**0.6879**	**0.7081**	**0.7183**	**0.7176**
$I \to IT$	CCQ (ours)	0.5074	0.5411	0.5414	0.5441	0.2512	0.2513	0.2529	0.2587	0.7063	0.6894	0.6989	0.6996
$T \to I$	CVH [13]	**0.5598**	0.5217	0.5129	0.4875	0.2309	0.2219	0.2214	0.2350	0.5972	0.6032	0.5738	0.5794
	IMH [20]	0.4380	0.4582	0.4186	0.4051	0.2394	0.2227	0.2333	0.1896	0.5374	0.5536	0.5513	0.5583
	CorrAE [8]	0.4303	0.4501	0.4634	0.4880	0.2688	0.2928	0.3478	0.3566	0.6142	0.6198	0.6247	0.6431
	CMSSH [5]	0.3783	0.3499	0.3944	0.4015	0.2926	0.2991	0.2537	0.2582	0.5868	0.5732	0.6176	0.6323
	SCM [28]	0.4449	0.4859	0.5105	0.5259	0.3157	0.3698	0.4239	0.4369	0.6037	0.5998	0.5805	0.6078
	QCH [26]	0.5020	0.5195	**0.5489**	**0.5622**	0.3426	0.3753	**0.4411**	**0.4565**	0.6258	0.6425	0.6485	0.6528
	CCQ (ours)	0.5359	**0.5410**	**0.5413**	0.5556	**0.3885**	**0.4000**	0.4222	0.4178	**0.6548**	**0.7026**	**0.7165**	**0.7266**
$T \to IT$	CCQ (ours)	0.6022	0.6925	0.7131	0.7153	0.6355	0.6351	0.6394	0.6405	0.6942	0.7151	0.7190	0.7416

via composite quantization in isomorphic space can learn balanced binary codes, hence its performance improves with longer codes.

It is interesting to observe that the performances of cross-modal search task $I \to T$ is generally better than that of intra-modal search task $I \to I$, while this observation does not hold for the counterparts $T \to I$ and $T \to T$. This seems abnormal at first sight as cross-modal search tasks are often more challenging than intra-modal search tasks due to semantic gap [18]. However, in general, text retrieval is much easier than image retrieval, making different modalities of the objects contribute differently the cross-modal retrieval performance. We believe that $T \to T$ is much easier than $T \to I$, but $I \to T$ may be easier than $I \to I$ because image-to-image retrieval is often the most difficult task. In the case of cross-modal task $I \to T$, the knowledge of text modality is transferred to image modality, making cross-modal retrieval easier. This shows cross-modal retrieval can be improved by knowledge transfer.

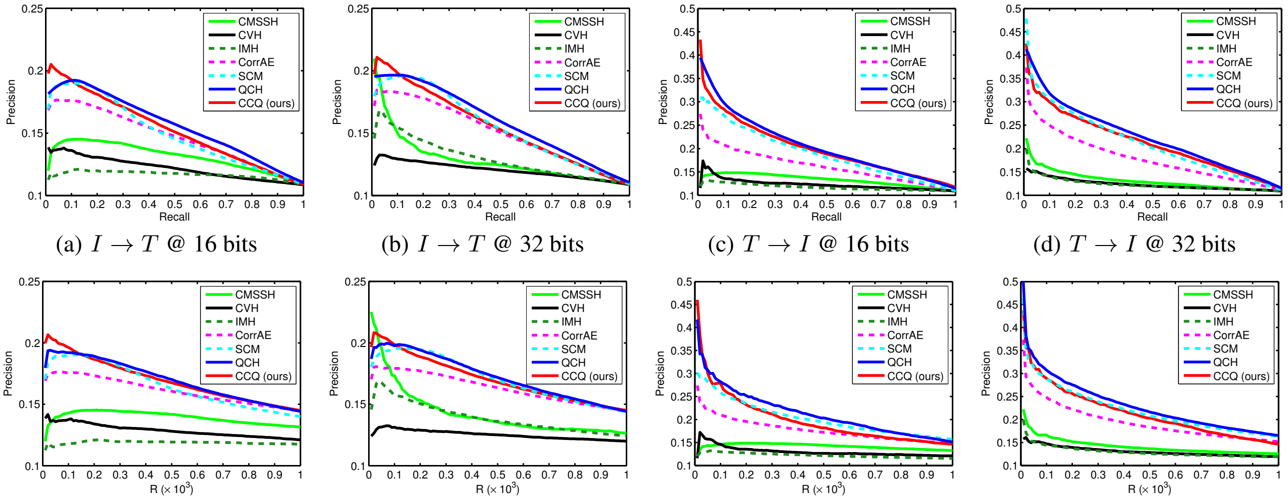
The precision-recall curves and the precision@top-R curves [34, 24] are illustrated in Figure 2. For space limitation, only the results of cross-modal tasks $I \to T$ and $T \to I$ are presented, while similar trends of results are observed on intra-modal tasks $I \to I$ and $T \to T$. CCQ shows the best cross-modal retrieval performance on all recall levels and top-R ranks. This validates that CCQ is capable for diverse retrieval scenarios, which may emphasize higher precision at smaller number of top-R retrieved results, i.e. Web search, or higher recall tolerating fairly lower precision, i.e. vertical search.

5.4.2 Results on Wiki

Table 2 also compares the search performance of CCQ and the state of the art methods on the Wiki dataset, which shows that CCQ significantly outperforms the unsupervised hashing methods CVH, IMH, and CorrAE, and performs comparably to supervised hashing methods SCM and QCH. A notable observation is that the MAPs are much smaller than those on the NUS-WIDE dataset. This is reasonable as the images of Wiki are of low-quality (low-resolution) and high-diversity, i.e. the text can well describe the semantics of the image-text pair while the image may not be well related to the semantics of the image-text pair, which makes it more challenging to capture the semantic correlations between image query and text database. Note that the texts of Wiki are featured articles which are well edited by experts and rich in semantic information, hence it is fairly easy to correlate a text query with the multimodal database.

The precision-recall curves and the precision@top-R curves [34, 24] are demonstrated in Figure 3. Again, CCQ is among the top-performing methods on all recall levels and all top-R ranks. A noticeable performance drop can be examined from the precision-recall curves to the precision@top-R curves. And this is because the Wiki dataset is very small-scale with only 2,173 database items, hence all relevant results will be retrieved at small R and no more relevant results can be further retrieved when R grows too large. This highlights the importance of evaluation with different metrics.

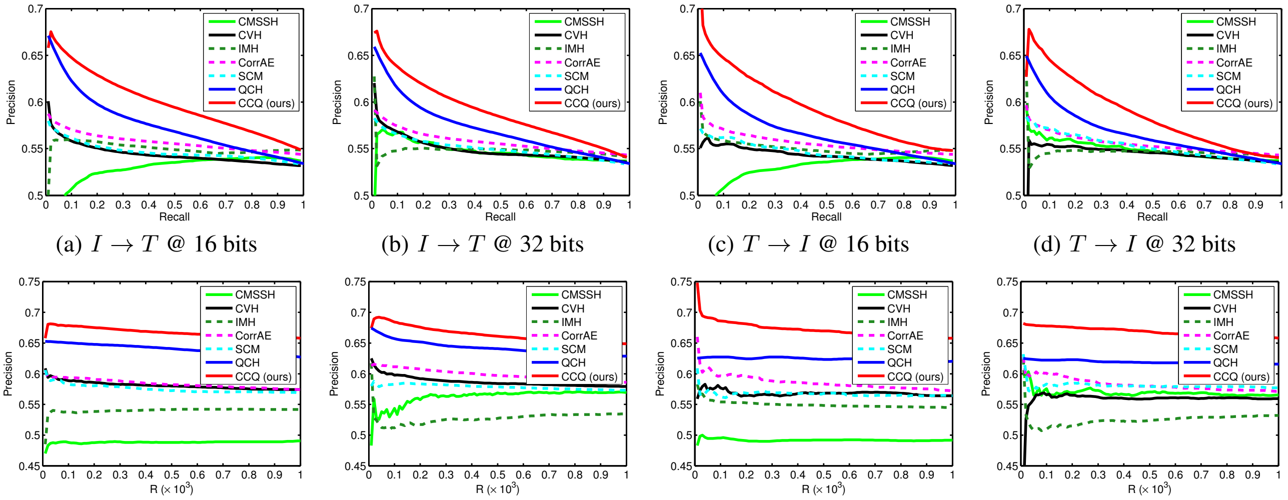
A crucial superiority of CCQ over the comparison methods lies in that CCQ jointly learns the isomorphic latent space and compact binary codes by minimizing both correlation and quantization errors in a unified optimization framework, while comparison methods merely learn the isomorphic space and binary codes in a separated two-step pipeline. As examined by CorrAE [8], the quality of searching with binary codes using Hamming distance is evidently inferior to searching with continuous features using Euclidean distance, due to substantial information loss by converting continuous features to binary codes without minimizing the quantization error. The search quality loss due to binarization is shown in Figure 5(a), and for CCQ, we use $\mathbf{R}^{v\mathsf{T}}\mathbf{x}_n^v$ for continuous features and $\mathbf{C}\mathbf{b}_n^v$ for



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

(e) $I \to T$ @ 16 bits (f) $I \to T$ @ 32 bits (g) $T \to I$ @ 16 bits (h) $T \to I$ @ 32 bits

Figure 3: Precision-recall curves (top) and precision@R curves (bottom) on Wiki cross-modal search tasks @ 16 and 32 bits.



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

(e) $I \to T$ @ 16 bits (f) $I \to T$ @ 32 bits (g) $T \to I$ @ 16 bits (h) $T \to I$ @ 32 bits

Figure 4: Precision-recall curves (top) and precision@R curves (bottom) on Flickr1M cross-modal search tasks @ 16 and 32 bits.

binary codes. We see that IMH and CorrAE suffer from substantial MAP loss (similar trends are observed from other methods) while CCQ is almost lossless to binarization. In other words, by jointly minimizing the correlation error and quantization error, CCQ can circumvent information loss and learn more accurate binary codes.

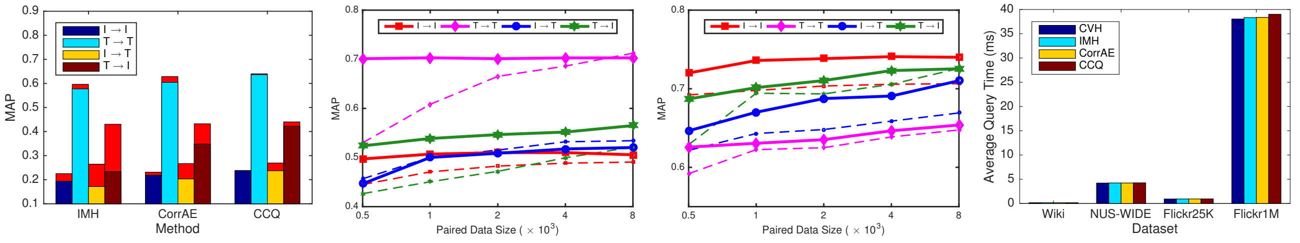
5.4.3 Results on Flickr1M

In practical retrieval systems, it is crucial to process large-scale datasets in both training and testing phases, and thus we compare CCQ with state of the art methods on large-scale Flickr1M dataset. We report the MAP results in Table 2 and illustrate the detailed precision-recall curves and precision@top-R curves in Figure 4. As mentioned before, we randomly select 10,000 image-text pairs as training set to learn hash functions if it is computationally too demanding to train these methods on the complete Flickr1M dataset. We can observe that CCQ significantly outperforms the comparison methods on all retrieval tasks and performs better with longer codes. This validates the superiority of CCQ in processing large-scale datasets, as the experimental setting on Flickr1M is consistent with real-word system setting where a sufficiently accurate model needs to be derived on a sufficiently large training set. We will examine CCQ's ability to process real semi-paired data in the sequel.

5.5 Semi-Paired Data Quantization

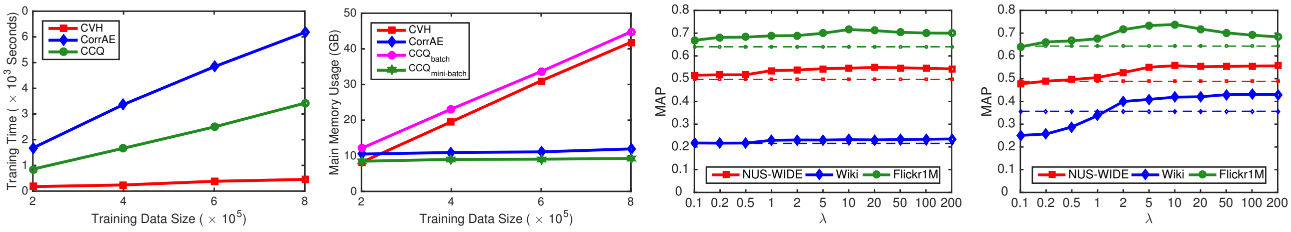
Most of the existing methods, including the ones in comparison, require full correspondences between different modalities. In other words, the multimodal data objects are fully paired, e.g. image-text pairs. As a result, these methods are unable to tackle more realistic scenarios in which only a limited number of paired data points are available. CCQ explores the idea of semi-supervised learning and can leverage both paired data (all modalities of the objects are available) and unpaired data (partial modalities of the objects are available) to boost the search quality when paired data are limited. To verify this, we consider the NUS-WIDE and Flickr1M datasets and for each dataset, we randomly sample as the training set 1) 10,000 images without text modality, 2) 10,000 texts without image modality, and 3) different numbers, i.e. $[0.5, 1, 2, 4, 8] \times 10^3$, of image-text pairs. We train CCQ with these semi-paired data and evaluate the search performance in terms of MAP @ 32 bits.

The search performances of CCQ on NUS-WIDE and Flickr1M



(a) MAP Loss (b) NUS-WIDE (c) Flickr1M (d) Search Efficiency

Figure 5: Effectiveness and efficiency experiments: (1) Loss of search quality in MAP (by red bars) due to conversion from continuous features to binary codes on Wiki. (b)–(c) the MAP of CCQ w.r.t. different numbers of paired data points (the number of unpaired data points is fixed to $10,000$). Solid lines indicate training with both paired and unpaired data, and dashed lines indicate training with only paired data. (d) Average search time (ms) for each query via lookup tables on Wiki, NUS-WIDE, Flickr25K, and Flickr1M.



(a) Time (b) Memory (c) $I \to T$ @ 32 bits (d) $T \to I$ @ 32 bits

Figure 6: Efficiency verification experiments: (a)–(b) Training time and memory costs of different methods on the complete Flickr1M dataset. CCQ with batch (mini-batch) training scales linearly (constantly) to the sample size. (c)–(d) The MAP of CCQ @ 32 bits versus parameter $\lambda \in [0.1, 200]$ for cross-modal retrieval tasks $I \to T$ and $T \to I$ on the NUS-WIDE, Wiki, and Flickr1M datasets.

are demonstrated in Figures 5(b) and 5(c) respectively, where solid lines indicate training with both paired and unpaired data, and dashed lines indicate training with only paired data. We can observe that when the number of paired data points is small, CCQ trained with both paired and unpaired data significantly outperforms CCQ trained with only paired data on most of the multimodal search tasks; when the number of paired data points increases, the search performance of CCQ will gradually saturate while the search quality of the two training paradigms will finally match. This clearly shows that CCQ can effectively leverage both paired and unpaired data (partial multimodal data) to boost search quality in a semi-paired data scenario.

An unexpected phenomenon is that semi-paired training slightly deteriorates search performance on task $I \to T$. We conjecture the plausible reason is that searching text database with image queries significantly relies on maximizing the image-text correlations to bridge the semantic gap between low-level image features and high-level image semantics, i.e. its associated texts. When the number of paired data points is obviously smaller than the number of unpaired data points, semi-paired training may tend to weaken correlation learning from image-text pairs and incur performance degradation.

5.6 Search Efficiency

To search for approximate nearest neighbors (ANN) in database for a given query, all methods in comparison perform linear scan using symmetric or asymmetric distance. Specifically, to compare a query vector with a database vector, CVH, IMH, and CorrAE all compute symmetric Hamming distance via lookup tables, and CCQ constructs a distance lookup table for each query that stores the Euclidean distances between the query and the multiple codebooks. As a result, CVH, IMH, CorrAE, and CCQ compute exactly the same number of table lookups for linear scan, while their costs of computing the query-codebook distance lookup tables are slightly different, which can be negligible as they are infinitesimal w.r.t. the cost of linear scan. For example, the cost of computing the distance lookup table for CCQ takes only less than 1% of the cost for linear scan on Flickr1M. The average search time of each query by CVH, IMH, CorrAE, and CCQ on the Wiki, NUS-WIDE, Flickr25K, and Flickr1M datasets is illustrated in Figure 5(d), from which we can observe that the search efficiency are comparable for all methods. While it is beyond the scope of this paper, we want to note that one can adopt a Multi-Index [2] approach to achieve sub-linear search complexity on the binary codes and further boost search efficiency.

5.7 Training Complexity

The training time and memory costs of CCQ scale linearly with the training sample size and hence can process large-scale dataset. To verify this, we follow [24] and use the complete Flickr1M dataset to evaluate the consumptions of training time and memory. CMSSH and IMH are not compared in this study since they require $O(N^2)$ complexity and run out of either time or memory on this dataset.

The comparison of training time costs is illustrated in Figure 6(a). We can observe that the training time of CCQ increases linearly with respect to the sample size. Due to multiple iterations between three sets of variables, i.e. transformation matrices $\mathbf{R}^v$, quantizer codebook $\mathbf{C}$, and modal-specific binary codes $\mathbf{B}^v$, CCQ is not as efficient as CVH. However, CCQ performs much more efficiently in time than CorrAE, which is a deep learning based method solving a time-demanding non-convex nonlinear optimization problem.

The training memory consumptions are compared in Figure 6(b). Both batch and mini-batch (large-scale) implementations of CCQ store the model parameters in memory, which are independent of training dataset size. For the batch implementation, all training data is loaded in memory, while for the mini-batch implementation, the training data is partitioned into multiple mini-batches while only one mini-batch is loaded in memory each time. Hence in the mini-batch (large-scale) implementation, the memory cost stays constant when training dataset size increases. We can flexibly allocate memory to each mini-batch to trade off memory and disk reading costs.

5.8 Parameter Sensitivity

Towards unsupervised multimodal retrieval, CCQ is designed to involve only two parameters, dimension of modality-isomorphic subspace D and modality trade-off weight λ, and the performance is expected to be stable against parameter variations. Since we have fixed $D = \min(\{P_v\}_{v=1}^{V}, H)$, we only inspect the sensitivity of λ.

We compute MAP @ 32 bits on both cross-modal retrieval tasks by varying λ between 0.1 and 200. The performance of CCQ w.r.t. parameter λ is shown in Figure 6(c) and 6(d). We see that CCQ can consistently outperform all the unsupervised baseline methods by a large margin with λ varying between 1 and 200. This validates that CCQ is robust against parameter selection and is applicable to unsupervised multimodal retrieval with easily-configured parameters.

6. CONCLUSION AND FUTURE WORK

In this paper, we have formally approached seamless multimodal hashing through a novel composite correlation quantization (CCQ). It integrates multimodal correlation and composite quantization into a seamless latent semantic analysis (LSA) framework, which yields compact binary codes that encode both intra-modal similarity and inter-modal correlation. The sharing of codebooks and binary codes across modalities enables joint learning of latent semantics that are maximally correlated in the isomorphic feature space, which serves as the key contributor to the efficacy of the proposed CCQ method.

In the future, we plan to equip our model with a deep learning architecture which can learn highly abstract nonlinear representations to better distill the correlation structures across multiple modalities.

7. ACKNOWLEDGMENTS

The authors would like to thank Dr Jingdong Wang for insightful comments. This work was supported by National Natural Science Foundation of China (61325008, 61502265), China Postdoctoral Science Foundation (2015T80088), National Science&Technology Supporting Program (2015BAH14F02), NSF grant III-1526499, and Tsinghua TNList Lab Fund for Big Data Science and Technology.

8. REFERENCES

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*. IEEE, 2006.

[2] A. Babenko and V. Lempitsky. The inverted multi-index. In *CVPR*, pages 3069–3076. IEEE, 2012.

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

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

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

[6] 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 *CIVR*. ACM, 2009.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[8] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *MM*. ACM, 2014.

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

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

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

[12] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*, 33(1):117–128, Jan 2011.

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

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

[15] 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 *SIGIR*. ACM, 2013.

[16] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.

[17] M. Norouzi and D. J. Fleet. Cartesian k-means. In *CVPR*. 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. *TPAMI*, 36(3):521–535, 2014.

[19] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[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 *SIGMOD*. ACM, 2013.

[21] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 15:2949–2980, 2014.

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

[23] Q. Wang, L. Si, and B. Shen. Learning to hash on partial multi-modal data. In *IJCAI*, pages 3904–3910, 2015.

[24] 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.

[25] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2009.

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

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

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

[29] D. Zhang, F. Wang, and L. Si. Composite hashing with multiple information sources. In *SIGIR*. ACM, 2011.

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

[31] F. Zhao, Y. Huang, L. Wang, and T. Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, 2015.

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

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

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