Semi-supervised Deep Quantization for Cross-modal Search

Xin Wang
xin_wang@tsinghua.edu.cn
Tsinghua University

Wenwu Zhu
wwzhu@tsinghua.edu.cn
Tsinghua University

Chenghao Liu
chliu@smu.edu.sg
Singapore Management University

ABSTRACT

The problem of cross-modal similarity search, which aims at making efficient and accurate queries across multiple domains, has become a significant and important research topic. Composite quantization, a compact coding solution superior to hashing techniques, has shown its effectiveness for similarity search. However, most existing works utilizing composite quantization to search multi-domain content only consider either *pairwise similarity* information or *class label* information across different domains, which fails to tackle the semi-supervised problem in composite quantization. In this paper, we address the semi-supervised quantization problem by considering: (i) *pairwise similarity* information (without class label information) across different domains, which captures the intra-document relation, (ii) cross-domain data with *class label* which can help capture inter-document relation, and (iii) cross-domain data with *neither pairwise similarity nor class label* which enables the full use of abundant unlabelled information. To the best of our knowledge, we are the first to consider both supervised information (*pairwise similarity + class label*) and unsupervised information (*neither pairwise similarity nor class label*) simultaneously in composite quantization. A challenging problem arises: how can we jointly handle these three sorts of information across multiple domains in an efficient way? To tackle this challenge, we propose a novel semi-supervised deep quantization (SSDQ) model that takes both supervised and unsupervised information into account. The proposed SSDQ model is capable of incorporating the above three kinds of information into one single framework when utilizing composite quantization for accurate and efficient queries across different domains. More specifically, we employ a modified deep autoencoder for better latent representation and formulate pairwise similarity loss, supervised quantization loss as well as unsupervised distribution match loss to handle all three types of information. The extensive experiments demonstrate the significant improvement of SSDQ over several state-of-the-art methods on various datasets.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

ACM Reference Format:
Xin Wang, Wenwu Zhu, and Chenghao Liu. 2019. Semi-supervised Deep Quantization for Cross-modal Search. In *Proceedings of the 27th ACM Int'l Conf. on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343031.3350934

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350934

1 INTRODUCTION

Compact coding, a commonly used solution to similarity search that considers both accuracy and efficiency simultaneously, has been widely explored in the past decade. Compact coding [24] achieves the goal of accurate and efficient search in databases through converting each data point into a compact representation with short codes. Typical solutions including hashing [15, 16, 27, 29, 37, 38, 40, 43] and quantization [8, 10, 25] have been extensively investigated by researchers in order to solve the similarity search problem within a single domain, while less works have been done for cross-modal retrieval across multiple domains. Among these compact coding solutions, composite quantization [48], as an enhanced variant of quantization, has been proved to outperform its hashing competitors.

However, existing works [3, 17, 44, 49] employing composite quantization to search multi-domain (cross-modal) data fails to handle the semi-supervised quantization problem because they either consider pairwise similarity information or (and) labelled information across different domains and thus ignore other additional information such as unlabelled information. Pairwise similarity information normally refers to the paired cross-domain (modal) data such that an image and a set of texts/tags describing it are able to form a pair, which can only capture the intra-document relation. Labelled information refers to cross-domain (modal) data with class label in addition to its pairwise similarity relation and unlabelled information refers to cross-domain (modal) data with neither pairwise similarity information nor class label information. In cross-modal similarity search, labelled information can be very useful in discovering the inter-document similarities and bridging the gaps between objects from different domains because class label information is able to explicitly indicate categories that objects belong to. Unlabelled information, though more likely being paid less attention compared to labelled information, is more pervasive in our daily life and can help enhance the robustness of machine learning models towards noises [53]. As such, all existing literature on quantization can only capture the intra-document relation, failing to make use of the labelled and unlabelled information, two types of helpful resources, for cross-domain (modal) information retrieval.

In this work, we solve the semi-supervised quantization problem through considering three kinds of information (i.e., paired, labelled and unlabelled) simultaneously. The motivation is very intuitive that incorporating additional useful information from labelled and unlabelled objects into our model can with no doubt help boost the search performance. However, these three types of information are organized in different formats: paired information needs the paired relationships between two particular objects in different domains;

labelled information requires a label for each object in different domains to indicate the category it belongs to; unlabelled information even does not provide any guidance for similarity search. Therefore, all existing works on composite quantization can not solve the above problem and it is very challenging to jointly take care of all three sorts of information across different domains in an efficient way. To overcome this challenge, we propose a novel **S**emi-**S**upervised **D**eep **Q**uantization (SSDQ) model that takes both supervised and unsupervised information into consideration. Our proposed SSDQ model is capable of incorporating the above three kinds of information into one single framework and utilizing composite quantization for accurate and efficient cross-modal queries at the same time. To be more concrete, we first employ a modified deep autoencoder for each domain to obtain a better latent representation of the input data given the recent success of deep neural networks in computer vision and multimedia. We then formulate supervised quantization loss (solvable through an iterative optimization process) to handle labelled information, formulate pairwise similarity loss through *maximum a posterior* (MAP) to handle paired information and formulate distribution match loss through *maximum mean discrepancy* (MMD) [28] to handle unlabelled information. We finally aggregate these three losses together to form the overall cross-modal retrieval loss and place it on the top level representation of the deep model so that the deep network parameters can be optimized through standard back-propagation (BP). The proposed SSDQ model is optimized through *Block Coordinate Descent* (BCD) [33] by alternatively learning quantization parameters (with deep model parameters fixed) and deep model parameters (with quantization parameters fixed). Experiments on two real-world datasets demonstrate the significant improvement of SSDQ over other state-of-the-art methods.

2 RELATED WORK

Cross-modal Hashing.

There exist some works such as composite hashing and effective multiple feature hashing [31, 47] examining the multi-modal representations, which though have a close relation to cross-modal hashing, are not specifically designed for it.

As a popular compact coding solution, cross-modal hashing normally maps data from different domains (modalities) into some common space (e.g., Hamming space) to ensure the comparability of the hash codes of multi-modality data in the new space. Objects from different domains may share one unified hash code or possess their own hash codes separately in the new space. In addition to designing a good hash function, cross-modal hashing models [2, 4, 5, 7, 9, 11–14, 21, 22, 26, 32, 36, 41, 42, 45, 46, 50–52, 54] focus on efficiently bridging the gaps between different domains for fast and accurate similarity search across multiple domains.

Cross-modal Quantization.

Quantization [8, 10], as a fairly new compact coding solution, has been developed for cross-modal similarity search in recent years [41]. Composite quantization, an improved version of quantization method shown to be superior to traditional hashing solutions [39, 48], has also been applied to handle effiecient and accurate cross-modal similarity search [3, 17, 49].

In particular, collaborative quantization [49] maps objects from different domains (such as images and texts) into a common space (not necessarily Hamming space) so that the quantized representations of similar objects from different domains can be forced to align with each other. Composite correlation quantization [17] adopts the similar quantization technique as collaborative quantization but instead uses a different similarity metric, Asymmetric Quantizer Distance (AQD), to calculate the distance between two objects. Collective deep quantization [3], on the other hand, resorts to two deep neural networks (convolutional neural network for image domains and multilayer perceptrons for text domain) and combine them together with the quantization techniques. Shared predictive deep quantization [44] assumes that the latent representation space consists of a shared subspace and two private subspaces where the shared components and the private components are captured separately. However, none of these works consider paired similarity information, labelled information, unlabelled information simultaneously and incorporate them into one unified framework to address the semi-supervised quantization problem.

3 SEMI-SUPERVISED DEEP QUANTIZATION

In this section, we first formally formulate the problem of cross-modal similarity search and then give a detailed description on our proposed semi-supervised deep quantization (SSDQ) model.

3.1 Problem Formulation

As a common setting in cross-modal similarity search [48], we assume that a database consists of data from two modalities/domains. For ease of understanding, we take images and texts as an example of two modalities. In image domain, we are given a set of labelled images $\{X^L, \boldsymbol{\ell}^X\} = \{(\boldsymbol{x}_i^L, \ell_i^X)\}_{i=1}^{n_X^L}$ and unlabelled images $X^{uL} = \{\boldsymbol{x}_i^{uL}\}_{i=1}^{n_X^{uL}}$. Similarly in text domain, we have a set of labelled texts $\{Y^L, \boldsymbol{\ell}^Y\} = \{(\boldsymbol{y}_i^L, \ell_i^Y)\}_{i=1}^{n_Y^L}$ and unlabelled texts $Y^{uL} = \{\boldsymbol{y}_i^{uL}\}_{i=1}^{n_Y^{uL}}$. In addition, there are also a set of paired objects across two domains (such as an image and its text description) $\{X^p, Y^p\} = \{(\boldsymbol{x}_i^p, \boldsymbol{y}_i^p)\}_{i=1}^{n^p}$. Here $\boldsymbol{x}_i^L, \boldsymbol{x}_i^{uL}, \boldsymbol{x}_i^p \in \mathcal{R}^{1\times d_X}$ and $\boldsymbol{y}_i^L, \boldsymbol{y}_i^{uL}, \boldsymbol{y}_i^p \in \mathcal{R}^{1\times d_Y}$ denote the feature representation of i-th labelled, unlabelled, paired image ($\boldsymbol{x}$) and text ($\boldsymbol{y}$) respectively, with d_X and d_Y as the feature dimension. Besides, we use superscript L, uL and p to denote corresponding labelled, unlabelled and paired data, thus n_X^L, n_X^{uL} and n^p denote the number of labelled, unlabelled and paired image objects (similar for text objects). For labelled data, $\ell_i^X, \ell_i^Y \in \{0, 1\}^k$, where the 1-value entry indicates the class label of $\boldsymbol{x}_i^L$, $\boldsymbol{y}_i^L$ and k is the number of categories for images/texts. For succinctness, we will ignore the corresponding superscript (i.e., L, uL, p, etc.) in the remaining of this paper when the context is clear. Our goal is that given an image (or text) query $\boldsymbol{x}$ (or $\boldsymbol{y}$), find the closest match sharing the same class label with the query in the text (or image) domain.

3.2 The Proposed SSDQ Model

It is challenging to simultaneously take care of paired, labelled and unlabelled data within one framework. Previous quantization works [3, 17, 49] on cross-modal similarity search only focus on

the paired data and fail to efficiently deal with all three kinds of data at the same time. To solve this challenge, we propose SSDQ, Semi-supervised Deep Quantization, to handle paired, labelled and unlabelled data within one framework. Given the success of deep representation in computer vision and natural language processing, we resort to a modified deep neural network structure to achieve a better feature representation for both image and text domain. The proposed SSDQ model first employs a deep autoencoder for initialization, then introduces supervised quantization loss for labelled data, pairwise similarity loss for paired data and unsupervised distribution match loss for unlabelled data. Figure 1 gives a detailed description of the proposed SSDQ model.

Model Initialization.

Deep autoencoder (DAE), as a deep architecture capable of reconstructing its input data and capturing the data manifolds smoothly [27] , has been widely used for the model initialization of deep neural network. Therefore, we follow the existing literature and initialize our model by resorting to deep autoencoder whose loss function is as follows.

$$\mathcal{L}_{x,DAE} = \sum_{i=1}^{n_x} \left\| \hat{\boldsymbol{x}}_i - \boldsymbol{x}_i \right\| \quad \text{and} \quad \mathcal{L}_{y,DAE} = \sum_{i=1}^{n_y} \left\| \hat{\boldsymbol{y}}_i - \boldsymbol{y}_i \right\|, \tag{1}$$

where $\boldsymbol{x}$ (or $\boldsymbol{y}$) is the original input to DAE in image (or text) domain and $\hat{\boldsymbol{x}}$ (or $\hat{\boldsymbol{y}}$) is the output from DAE in image (or text) domain. In addition, Relu [23], i.e., g(x) = max(0,x), is adopted as the activation function to mitigate the problem of vanishing gradient. We remark that other deep model related common settings are the same as literature and therefore will be omitted in this paper due to the page limit.

Supervised Quantization Loss.

For the labelled objects from image domain $\{X, \boldsymbol{\ell}^X\} = \{(\boldsymbol{x}_i, \ell_i^X)\}_{i=1}^{n_X^L}$ and text domain $\{Y, \boldsymbol{\ell}^Y\} = \{(\boldsymbol{y}_j, \ell_j^Y)\}_{j=1}^{n_Y^L}$, we develop a novel supervised deep quantization strategy that supports fast and efficient cross-modal similarity search through approximating the deep representation of each single database object (image or text) with a vector composed from a dictionary of base items whose indices will then form the short code representing this database object.

For objects in text database, we are given a dictionary set $\boldsymbol{D}$ consisting of M dictionaries $\{\boldsymbol{D}_m\}_{m=1}^M$ and $\boldsymbol{D}_m = [\boldsymbol{d}_{m,1}, \boldsymbol{d}_{m,2}, \ldots, \boldsymbol{d}_{m,K-1}, \boldsymbol{d}_{m,K}]$, where K is the number of base items in each dictionary $\boldsymbol{D}_m$. We employ the methodology of composite quantization [48] to approximate the top level deep representation of each text object $\boldsymbol{z}$ by summing up M base items (each of which is chosen from only one of the K base items in $\boldsymbol{D}_m$), i.e., $\boldsymbol{z} = \sum_{m=1}^M \boldsymbol{d}_{m,k_m}$. Thus, $\boldsymbol{z}$ can be encoded by a short code $(k_1, k_2, \ldots, k_M)$ which may take much less space than storing $\boldsymbol{z}$ directly. Similarly for objects in image database, we use another dictionary set $\boldsymbol{C} = \{\boldsymbol{C}_m\}_{m=1}^M$ where $\boldsymbol{C}_m = [\boldsymbol{c}_{m,1}, \boldsymbol{c}_{m,2}, \ldots, \boldsymbol{c}_{m,K-1}, \boldsymbol{c}_{m,K}]$ to encode their top level deep representations.

On the one hand, we require the quantization approximation to be close to the encoded deep representation, i.e.,

$$\min_{\boldsymbol{C}} \left\| \sum_{m=1}^{M} \boldsymbol{c}_{m,k_m} - \boldsymbol{z}_i \right\|_2^2 \quad \text{and} \quad \min_{\boldsymbol{D}} \left\| \sum_{m=1}^{M} \boldsymbol{d}_{m,k_m} - \boldsymbol{z}_j \right\|_2^2, \tag{2}$$

where $\boldsymbol{z}_i, \boldsymbol{z}_j$ are top level deep representations for image objects and text objects, respectively. On the other hand, we also need the label information to guide inter-document similarity learning across different domains. Assuming the label contains totally $\mathbb{C}$ classes and the dimensions of top level deep representations for objects in different domains (databases) are r, we would like to predict the class labels given the top level deep representations (or their corresponding compact codes) of the objects in different databases, which can be formulated in a regression form as follows:

$$\min_{\boldsymbol{W}_X} \sum_{i=1}^{n_X^L} \left\| \ell_i^X - \boldsymbol{W}_X^T \boldsymbol{z}_i \right\|_2^2 \quad \text{and} \quad \min_{\boldsymbol{W}_Y} \sum_{j=1}^{n_Y^L} \left\| \ell_j^Y - \boldsymbol{W}_Y^T \boldsymbol{z}_j \right\|_2^2, \tag{3}$$

where $\boldsymbol{W}_X \in \mathcal{R}^{r\times\mathbb{C}}$ and $\boldsymbol{W}_Y \in \mathcal{R}^{r\times\mathbb{C}}$ are classification matrices to assign each object to its predicted class (label). Let $\boldsymbol{CB}_i$ and $\boldsymbol{DF}_j$ be the vector representations of $\sum_{m=1}^M \boldsymbol{c}_{m,k_m}$ and $\sum_{m=1}^M \boldsymbol{d}_{m,k_m}$ with $\boldsymbol{B}_i = \left[\boldsymbol{b}_{i,1}^T, \boldsymbol{b}_{i,2}^T, \ldots, \boldsymbol{b}_{i,M-1}^T, \boldsymbol{b}_{i,M}^T\right]^T$ and $\boldsymbol{F}_j = \left[\boldsymbol{f}_{j,1}^T, \boldsymbol{f}_{j,2}^T, \ldots, \boldsymbol{f}_{j,M-1}^T, \boldsymbol{f}_{j,M}^T\right]^T$, we require $\boldsymbol{b}_{i,m}^T$ and $\boldsymbol{f}_{j,m}^T$ to be vectors of K dimensions with the constraint that only one entry has value 1 and the values of remaining $K-1$ entries are all 0. Thus, we are able to obtain the overall loss function for the deep supervised quantization by putting Eq (2) and Eq (3) together:

$$\begin{aligned}
\mathcal{L}_{quant} = \min_{\Theta} & \sum_{i=1}^{n_X^L} \left\| \ell_i^X - \boldsymbol{W}_X^T \boldsymbol{C}\boldsymbol{B}_i \right\|_2^2 + \sum_{j=1}^{n_Y^L} \left\| \ell_j^Y - \boldsymbol{W}_Y^T \boldsymbol{D}\boldsymbol{F}_j \right\|_2^2 \\
& + \eta \left(\sum_{i=1}^{n_X^L} \left\| \boldsymbol{C}\boldsymbol{B}_i - \boldsymbol{z}_i \right\|_2^2 + \sum_{j=1}^{n_Y^L} \left\| \boldsymbol{D}\boldsymbol{F}_j - \boldsymbol{z}_j \right\|_2^2 \right) \\
& + \tau \left(\left\| \boldsymbol{W}_X \right\|_2^2 + \left\| \boldsymbol{W}_Y \right\|_2^2 \right) \\
\text{s.t.} \quad & \sum_{p=1}^{M} \sum_{q=1, q\neq p}^{M} \boldsymbol{b}_{i,p}^T \boldsymbol{C}_p^T \boldsymbol{C}_q \boldsymbol{b}_{i,q} = \epsilon_1 \\
& \sum_{p=1}^{M} \sum_{q=1, q\neq p}^{M} \boldsymbol{f}_{j,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{j,q} = \epsilon_2,
\end{aligned} \tag{4}$$

where $\Theta = \left\{\boldsymbol{W}_X, \boldsymbol{W}_Y, \boldsymbol{C}, \boldsymbol{D}, \{\boldsymbol{B}_i\}_{i=1}^{n_X^L}, \{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}, \epsilon_1, \epsilon_2\right\}$ and η, τ control the quantization term, regularization term respectively. We remark that ϵ_1 and ϵ_2 are used to constrain $\sum_{p=1}^M \sum_{q=1,q\neq p}^M \boldsymbol{b}_{i,p}^T \boldsymbol{C}_p^T \boldsymbol{C}_q \boldsymbol{b}_{i,q}$ and $\sum_{p=1}^M \sum_{q=1,q\neq p}^M \boldsymbol{f}_{j,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{j,q}$ to be constants, which is first introduced by Wang et al. [48] as *constant inter-dictionary-element product* to fast calculate the distance between a query and a database object (time complexity $O(M)$ for searching).

Pairwise Similarity Loss.

To utilize the pairwise information of paired objects $\{X, Y\} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n^P}$, we define $Ip(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1$ (short for *Is pair*) if object $\boldsymbol{x}_i$ from one domain and object $\boldsymbol{y}_j$ from another domain belong to one pair, and $Ip(\boldsymbol{x}_i, \boldsymbol{y}_j) = 0$ otherwise. Thus, we define the probability of $Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)$ as follows:

$$\begin{aligned}
& \mathcal{P}\left(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)\right) \\
=& \delta(\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle)^{Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)} \left(1 - \delta(\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle)\right)^{1-Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)},
\end{aligned} \tag{5}$$

where $\boldsymbol{z}_i, \boldsymbol{z}_j$ are the top level representations of $\boldsymbol{x}_i, \boldsymbol{y}_j$ respectively and $\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle$ indicates the inner product of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, which is commonly adopted to measure the similarity of $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$. Besides, we denote $\delta(x)$ as the sigmoid function $\delta(x) = \frac{1}{1+\exp(-\alpha x)}$, where coefficient α lies in $[0, 1]$ to avoid the vanishing gradient problem during back-propagation. We note that a larger value of $\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle$ indicates a larger value for $\mathcal{P}(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1)$ (i.e., a higher

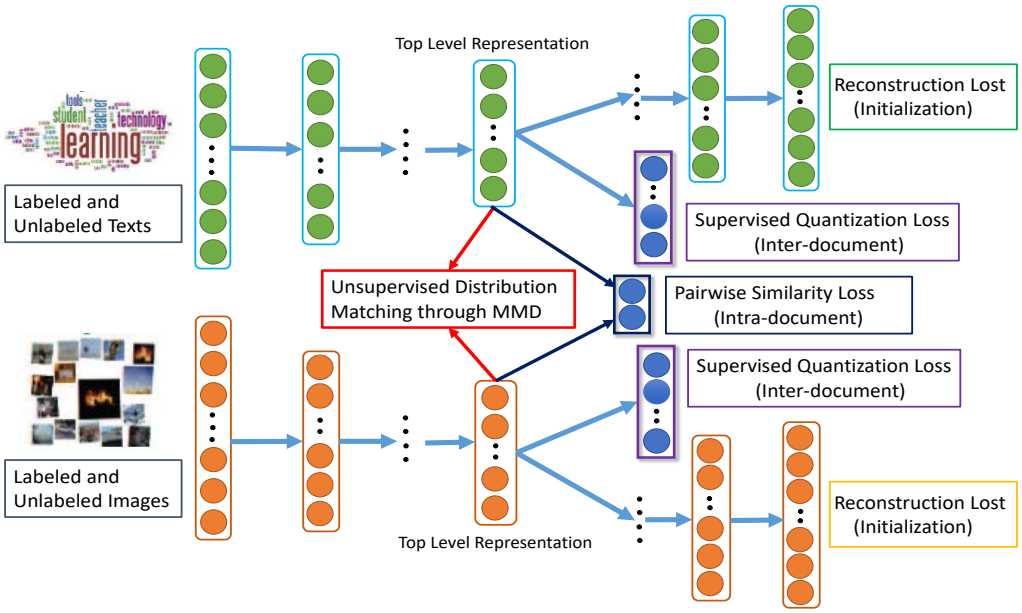


Figure 1: Framework of the proposed semi-supervised deep quantization (SSDQ) model

probability for $Ip(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1$) and vice versa. Our goal is to maximize $\prod_{i=1}^{n^p}\prod_{j=1}^{n^p}\mathcal{P}(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j))$, which in turn equals to minimize $-\ln\prod_{i=1}^{n^p}\prod_{j=1}^{n^p}\mathcal{P}(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j))$. The loss function $\mathcal{L}_{pair}$ for pairwise similarity is as follows:

$$\begin{aligned}\mathcal{L}_{pair} &= \sum_{i=1}^{n^p}\sum_{j=1}^{n^p}\Big(-\ln\mathcal{P}\big(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)\big)\Big)\\ &= \sum_{i=1}^{n^p}\sum_{j=1}^{n^p}-\Big(Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)\ln\frac{1}{1+\exp\big(-\alpha\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\big)}\\ &+\big(1-Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)\big)\ln\frac{\exp\big(-\alpha\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\big)}{1+\exp\big(-\alpha\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\big)}\Big)\\ &= \sum_{i=1}^{n^p}\sum_{j=1}^{n^p}\Big(\ln\Big(1+\exp\big(-\alpha\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\big)\Big)+\alpha\big(1-Ip(\boldsymbol{x}_i, \boldsymbol{y}_j)\big)\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\Big).\end{aligned} \tag{6}$$

Unsupervised Distribution Match Loss.

Besides labelled data and paired data, there is normally a larger amount of unlabelled data surrounding us, which may be a rich information source that can help further improve the accuracy and robustness [53] of cross-modal similarity search. This being the case, we will next focus on utilizing unlabelled objects from different databases, i.e., $X = \{\boldsymbol{x}_i\}_{i=1}^{n_X^{uL}}$ and $Y = \{\boldsymbol{y}_j\}_{j=1}^{n_Y^{uL}}$, for cross-modal similarity search. However, different from labelled or paired data that is capable of bridging different domains through offering supervised information such as the labels or corresponding pairs, it is more difficult to connect the unlabelled data from one domain with that from another domain, which poses a challenging problem to us. To solve this problem, we resort to *Maximum Mean Discrepancy* (MMD) [28] to make connections between top level deep representations of unlabelled objects in different databases. This is motivated by MMD's previous success in cross-domain adaptation and its capability of constructing regularization terms during feature representation learning so that the learned feature representations in different domains are constrained to be as identical as possible. Given a number of unlabelled objects from different databases, say image objects and text objects, we assume that the marginal distributions over the top level deep representations across these two domains should be similar. More concretely, we employ MMD as the distance measure for comparisons between these two distributions. Although it would be more precise, according to the strict definition of MMD, to first map top level deep representations to *Reproducing Kernel Hilbert Space* (RKHS) before the measurement, we follow existing literature [18] and omit the mapping procedure for the sake of simplicity. As such, the loss function $\mathcal{L}_{match}$ for unlabelled data across two domains through maximizing the mean distribution discrepancy is as follows:

$$\mathcal{L}_{match} = \left\| \frac{1}{n_X^{uL}}\sum_{i=1}^{n_X^{uL}}\boldsymbol{z}_i - \frac{1}{n_Y^{uL}}\sum_{j=1}^{n_Y^{uL}}\boldsymbol{z}_j \right\|_2^2, \tag{7}$$

where $\boldsymbol{z}_i, \boldsymbol{z}_j$ are the top level representations for image object ($\boldsymbol{x}_i$) and text object ($\boldsymbol{y}_j$) respectively.

Total Loss for Cross-modal Retrieval.

Putting the three loss functions for labelled, paired and unlabelled data all together, we come up with our complete loss function for cross-modal similarity search:

$$\mathcal{L}_{cross} = \alpha\mathcal{L}_{quant} + \beta\mathcal{L}_{pair} + \gamma\mathcal{L}_{match} + \lambda\mathcal{L}_{reg}, \tag{8}$$

where $\alpha, \beta, \gamma, \lambda$ control the relative importance of different loss terms and $\mathcal{L}_{reg}$ contains norm-2 regularizations of all the parameters including weights and biases in each layer of deep autoencoder for both domains. For succinctness, we denote the neural weights as well as biases in image domain as θ_X and those in text domain as θ_Y. We adopt the same way of dealing parameter regularizations as in state-of-the-art works on deep neural networks and will not go into details given the page limit. We would like to point out that placing different losses on top level representation of the SSDQ model is motivated by the claim from Srivastava and Salakhutdinov's work [34] that cross-modal data possesses more explicit

relationships in higher level space of deep neural networks, where more semantic information can be preserved.

Querying.

Given a new query $\boldsymbol{n}$ in image domain, we can easily obtain its top level deep representation $\boldsymbol{z}_n$. After the training procedure terminates, any data point $\boldsymbol{y}$ in the text database (same for image database) can be transformed by our deep structure to its corresponding top level representation $\boldsymbol{z}_y$ which in turn is approximated by its quantized form, i.e., $\boldsymbol{z}_y = \sum_{m=1}^{M} \boldsymbol{d}_{m,k_m}$. Thus the distance between $\boldsymbol{n}$ and any data point in the text database $\boldsymbol{y}$ can be approximately calculated as follows:

$$Dist(\boldsymbol{n}, \boldsymbol{y}) = \left\|\boldsymbol{z}_n - \sum_{m=1}^{M} \boldsymbol{d}_{m,k_m}\right\|_2^2$$
$$= \sum_{m=1}^{M} \left\|\boldsymbol{z}_n - \boldsymbol{d}_{m,k_m}\right\|_2^2 - (M-1)\|\boldsymbol{z}_n\|_2^2 + \sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{d}_{p,k_p}^T \boldsymbol{d}_{q,k_q}, \quad (9)$$

where in Eq (9)

$$\sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{d}_{p,k_p}^T \boldsymbol{d}_{q,k_q} = \sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{f}_{y,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{y,q} = \epsilon_2$$

equals to a constant ϵ_2 according to the second constraint in Eq (2) and $(M-1)\|\boldsymbol{z}_n\|_2^2$ in Eq (9) is also a constant for any database point (in vector form) when query $\boldsymbol{z}_n$ is given in advance. Therefore, we ignore these two constant terms when ranking the approximate distances between a new query and data points in the database as they have no impact on the ranking result. We first pre-compute a distance table containing the distances between query $\boldsymbol{z}_n$ and all the base items in every dictionary $\boldsymbol{D}_m \in \boldsymbol{D}$, which takes upto length MK (K base items in one dictionary times M dictionaries in total). Thus the calculation of $\sum_{m=1}^{M} \left\|\boldsymbol{z}_n - \boldsymbol{d}_{m,k_m}\right\|_2^2$, with the two constant terms in Eq (9) being ignored, will only take $O(M)$ lookups in the distance table plus $O(M)$ additions.

Searching image database given a new query in the text domain is symmetrically a similar procedure in essential.

4 OPTIMIZATION

After resorting to the **Model Initialization** procedure for an adequate initialization of SSDQ model,we utilize fine tuning to enhance the performance of cross-modal similarity search for SSDQ.

Our goal is to obtain the optimal parameters Θ in supervised quantization and $\theta = \{\theta_X, \theta_Y\}$ in deep structure through minimizing $\mathcal{L}_{cross}$. However, given the fact that optimizing Θ and θ simultaneously while keeping $\mathcal{L}_{cross}$ minimized is intractable and difficult, we adopt the *Block Coordinate Descent* (BCD) [33] optimization strategy to alternatively learn the parameters in an iterative way.

4.1 Optimizing Θ

We fix the deep structure parameters θ to learn Θ which consists of eight variables: $\boldsymbol{W}_X, \boldsymbol{W}_Y, \boldsymbol{C}, \boldsymbol{D}, \{\boldsymbol{B}_i\}_{i=1}^{n_x^L}, \{\boldsymbol{F}_j\}_{j=1}^{n_y^L}, \epsilon_1, \epsilon_2$. By rewriting Eq (4) with the two constraint terms incorporated into the loss function, we have:

$$\mathcal{J}^{quant} = \sum_{i=1}^{n_X^L} \left\|\ell_i^X - \boldsymbol{W}_X^T \boldsymbol{C}\boldsymbol{B}_i\right\|_2^2 + \sum_{j=1}^{n_Y^L} \left\|\ell_j^Y - \boldsymbol{W}_Y^T \boldsymbol{D}\boldsymbol{F}_j\right\|_2^2$$
$$+ \eta\left(\sum_{i=1}^{n_X^L} \left\|\boldsymbol{C}\boldsymbol{B}_i - \boldsymbol{z}_i\right\|_2^2 + \sum_{j=1}^{n_Y^L} \left\|\boldsymbol{D}\boldsymbol{F}_j - \boldsymbol{z}_j\right\|_2^2\right)$$
$$+ \mu\sum_{i=1}^{n_X^L}\left(\sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{b}_{i,p}^T \boldsymbol{C}_p^T \boldsymbol{C}_q \boldsymbol{b}_{i,q} - \epsilon_1\right)$$
$$+ \mu\sum_{j=1}^{n_Y^L}\left(\sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{f}_{j,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{j,q} - \epsilon_2\right)$$
$$+ \tau\left(\|\boldsymbol{W}_X\|_2^2 + \|\boldsymbol{W}_Y\|_2^2\right), \quad (10)$$

where μ is the penalty controlling parameter. We iteratively update each variable with the others in Θ fixed.

Learning $\boldsymbol{D}$. Assuming other variables including $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$ and $\boldsymbol{W}_Y$ are fixed, solving $\boldsymbol{D}$ becomes an unconstrained nonlinear optimization problem which can be tackled through the L-BFGS algorithm (**l**imited version of the **B**royden-**F**letcher-**G**oldfarb-**S**hanno algorithm) requiring Eq (10) as well as its partial derivative with respect to $\boldsymbol{D}_m$ as the input.

$$\frac{\partial \mathcal{J}^{quant}}{\partial \boldsymbol{D}_m} = \sum_{j=1}^{n_Y^L}\left(2\boldsymbol{W}_Y\left(\boldsymbol{W}_Y^T\boldsymbol{D}\boldsymbol{F}_j - \ell_j^Y\right)\boldsymbol{f}_{j,m}^T + 2\eta\left(\boldsymbol{D}\boldsymbol{F}_j - \boldsymbol{z}_j\right)\boldsymbol{f}_{j,m}^T\right.$$
$$\left. +4\mu\left(\sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{f}_{j,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{j,q} - \epsilon_2\right)\left(\sum_{g=1,g\neq m}^{M} \boldsymbol{D}_g \boldsymbol{f}_{j,g}\right)\boldsymbol{f}_{j,m}^T\right). \quad (11)$$

Learning $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$. Optimizing $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$ with other variables in Θ fixed can be decomposed into n_Y^L subproblems, each of which is NP-hard [48],

$$\mathcal{J}_j^{quant}(\boldsymbol{F}_j) = \left\|\ell_j^Y - \boldsymbol{W}_Y^T\boldsymbol{D}\boldsymbol{F}_j\right\|^2 + \eta\left\|\boldsymbol{D}\boldsymbol{F}_j - \boldsymbol{z}_j\right\|^2$$
$$+\mu\left(\sum_{p=1}^{M}\sum_{q=1,q\neq p}^{M} \boldsymbol{f}_{j,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{j,q} - \epsilon_2\right)^2. \quad (12)$$

In essential, Eq (12) is a high-order Markov Random Field problem which can be tackled through alternatively solving $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$ by employing the *Iterated Conditional Modes* (ICM) algorithm [1]. More concretely, we first locate the best base item in dictionary $\boldsymbol{D}_m$ that minimizes Eq (12) through exhaustively searching every base item in $\boldsymbol{D}_m$, assuming $\sum_{g=1,g\neq m}^{M} \boldsymbol{f}_{j,g}$ are all fixed. Then the corresponding entry of $\boldsymbol{f}_{j,g}$ is set to 1 and all other entries are set to 0. This procedure (ICM algorithm) is guaranteed to converge.

Learning $\boldsymbol{W}_Y$. By fixing $\boldsymbol{D}$ and $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$, the optimal solution of $\boldsymbol{W}_Y$ can be computed as follows:

$$\boldsymbol{W}_Y^* = \left(\boldsymbol{H}\boldsymbol{H}^T + \tau\boldsymbol{I}\right)^{-1}\boldsymbol{H}\boldsymbol{L}^T, \quad (13)$$

where $\boldsymbol{H} = \left[\boldsymbol{D}\boldsymbol{F}_1, \boldsymbol{D}\boldsymbol{F}_2, \ldots, \boldsymbol{D}\boldsymbol{F}_{n_Y^L-1}, \boldsymbol{D}\boldsymbol{F}_{n_Y^L}\right] \in \mathcal{R}^{r\times n_Y^L}$, $\boldsymbol{I}$ is a $r \times r$ identity matrix and $\boldsymbol{L} = \left[\ell_1^Y, \ell_2^Y, \ldots, \ell_{n_Y^L-1}^Y, \ell_{n_Y^L}^Y\right] \in \mathcal{R}^{\mathbb{C}\times n_Y^L}$. For ease of notation, here we extend scaler ℓ_j^Y to its vector form, e.g., assuming ℓ_j^Y indicates that $\boldsymbol{y}_j$ belongs to category $j \in [1, \mathbb{C}]$, then

we extend ℓ_j^Y to a $\mathbb{C} \times 1$ vector whose j-th entry is 1 and all other entries are 0.

Learning ϵ_2. Similarly as the learning of $\boldsymbol{W}_Y$, we can more easily calculate the solution of ϵ_2 when fixing $\boldsymbol{D}$ and $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$:

$$\epsilon_2^* = \frac{1}{n_Y^L} \sum_{j=1}^{n_Y^L} \sum_{p=1}^{M} \sum_{q=1, q\neq p}^{M} \boldsymbol{f}_{y,p}^T \boldsymbol{D}_p^T \boldsymbol{D}_q \boldsymbol{f}_{y,q} \tag{14}$$

Learning $\boldsymbol{C}$. Similar to the learning procedure of $\boldsymbol{D}$.

Learning $\{\boldsymbol{B}_j\}_{j=1}^{n_X^L}$. Similar to the learning procedure of $\{\boldsymbol{F}_j\}_{j=1}^{n_Y^L}$.

Learning $\boldsymbol{W}_X$. Similar to the learning procedure of $\boldsymbol{W}_Y$.

Learning ϵ_1. Similar to the learning procedure of ϵ_2.

4.2 Optimizing θ

Similarly, we fix the quantization parameters Θ to learn θ, which can be efficiently achieved by the well established back-propagation from the top layers down through the whole deep structure.

Algorithm 1 shows the details of our proposed SSDQ model.

Algorithm 1: Semi-supervised Deep Quantization

Data: $\{X^L, \boldsymbol{\ell}^X\} = \{(\boldsymbol{x}_i^L, \ell_i^X)\}_{i=1}^{n_X^L}$, $X^{uL} = \{\boldsymbol{x}_i^{uL}\}_{i=1}^{n_X^{uL}}$,
$\{Y^L, \boldsymbol{\ell}^Y\} = \{(\boldsymbol{y}_i^L, \ell_i^Y)\}_{i=1}^{n_Y^L}$, $Y^{uL} = \{\boldsymbol{y}_i^{uL}\}_{i=1}^{n_Y^{uL}}$,
$\{X^p, Y^p\} = \{(\boldsymbol{x}_i^p, \boldsymbol{y}_i^p)\}_{i=1}^{n^p}$

// Initialization

Initialize model parameters Θ and θ.

// Train SSDQ through Block Coordinate Descent strategy

repeat

  Update Θ according to Section 4.1, given top level representations $\{\boldsymbol{z}_i\}_{i=1}^{n_X}$ $\{\boldsymbol{z}_j\}_{j=1}^{n_Y}$ obtained through fixed deep neural network parameters θ.

  Calculate $\mathcal{L}_{cross}(\Theta; \theta)$ by Eq (8).

  $\theta' \leftarrow \theta - \rho * \frac{\partial \mathcal{L}_{cross}}{\partial \theta}$, where ρ is the learning rate.

  $\theta \leftarrow \theta'$

until *converge*

Result: Optimized Θ, θ

4.3 Complexity Analysis

The time complexity of deep neural network is exactly the same as state-of-the-art vanilla autoencoder based models. The complexity of quantization is : $O(M^2K^2d)$ for inner product tables, $O(NMKdT_c)$ for updating c (similar for updating d) where T_c is the iteration number, $O(NM^2)$ for updating ϵ and $O(NMdT_lT_c)$ for updating $\mathbf{C}_m$ (similar for $\mathbf{D}_m$) where T_l is the number of searches in L-BFGS.

5 EMPIRICAL EXPERIMENTS

In this section, we compare the performance of our proposed SSDQ model with several state-of-the-art approaches on two web image datasets, NUS-WIDE [6] and Flickr1M [20] whose detailed statistics as well as the corresponding parameter settings will be shown in Appendix A and B.

5.1 Experimental Settings

Comparative Methods. We compare our proposed SSDQ model with eight state-of-the-art methods including five hashing approaches, i.e., cross view hashing (**CVH**) [13], data fusion hashing (**CMSSH**) [2], semantic correlation maximization hashing (**SCM**) [46] semantics-preserving cross-view hashing (**SePH**) [14], deep cross-modal hashing (**DCMH**) [11], and three quantization approaches, i.e., cross-modal collaborative quantization (**CMCQ**) [49], collective deep quantization (**CDQ**) [3], shared predictive cross-modal deep quantization (**SPDQ**) [44]. For the sake of fair comparisons, we simply employ SIFT or VGG features as inputs for shallow models, and feed these features to a 5-layer deep structure for image domain and a 4-layer deep structure for text domain as inputs for deep models.

Evaluation Metrics. We evaluate all the algorithms in terms of two cross-modal search tasks: i) searching texts given images as queries and ii) searching images given texts as queries. We follow previous works [17, 49] and adopt two metrics, *Precision@T* and *MAP@T*, to evaluate the searching quality. *MAP@T* is defined as the average precision $AP_q@T$ over all queries:

$$AP_q@T = \frac{\sum_{t=1}^{T} Precision@t \cdot \delta(t)}{\sum_{t=1}^{T} \delta(t)}, \tag{15}$$

where t is the number of retrieved objects and $Precision@t$ is the precision of the first t returned objects. Besides, if the t_{th} returned object shares the same label with the query, then $\delta(t) = 1$, otherwise $\delta(t) = 0$.

5.2 Experimental Results

Following previous works [7, 17, 49], we report *MAP@50* for all night comparative methods with different lengths of binary codes ranging from 8 bits to 32 bits on NUS-WIDE (Table 1) and Flickr1M (Table 2). We note that the length of binary codes is calculated as follows: assuming we have M dictionaries each of which has K base items, then the length of binary codes is $M\log K$.

In general, we observe that using deep features with higher dimensions can produce more accurate results than using hand-crafted features (VGG16 features have 4092 dimensions while SIFT features only possess 500 or 3857 dimensions). The credits of better searching performances may go to the richer information lying in the deep representations obtained by pre-trained models.

Specifically, our proposed SSDQ model significantly outperforms other state-of-the-art methods on NUS-WIDE dataset and achieves the best search quality among all the nine comparative methods on Flickr dataset, with both SIFT and VGG16 features. Besides, we also observe from both NUS-WIDE and Flickr datasets that the performances of SSDQ get better as more bits are used, indicating that our proposed approach can utilize the extra code length more efficiently to improve search quality.

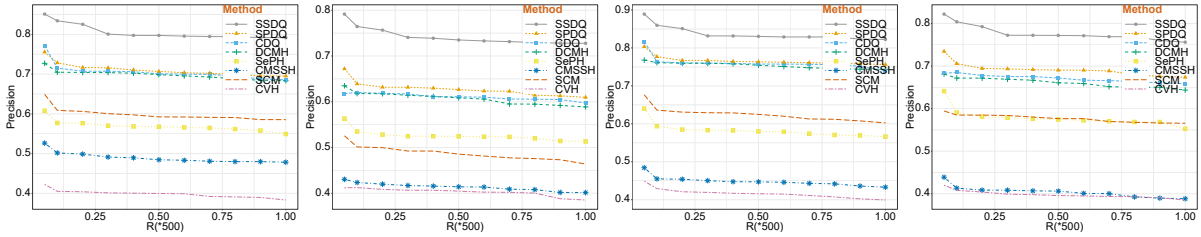
Precision v.s. Number of Top Retrieved Items. Figure 2 and Figure 3 present the Precision-#Retrieved Items curves on NUS-WIDE (SIFT and VGG16) and Flickr1M (SIFT and VGG16) with code length 32. The results with code length 16 demonstrate similar patterns and we will present them in supplementary file due to page limit. In each figure, subfigure (a) shows the results for the

Task	Method	NUS-WIDE (SIFT)				NUS-WIDE (VGG16)			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
Text to Image	CVH	0.3833	0.4009	0.4091	0.4042	0.4226	0.4255	0.4303	0.4246
	SCM	0.5470	0.5770	0.5918	0.6092	0.6066	0.6113	0.6191	0.6360
	CMSSH	0.3679	0.4256	0.4952	0.5009	0.4209	0.4415	0.4450	0.4529
	SePH	0.5209	0.5459	0.5834	0.5793	0.5810	0.5660	0.5981	0.5905
	CMCQ	0.5888	0.5951	0.6144	0.6193	0.6155	0.6296	0.6225	0.6969
	DCMH	0.5868	0.6258	0.6942	0.7110	0.5944	0.6790	0.7232	0.7662
	CDQ	0.6450	0.6519	0.7027	0.7183	0.6170	0.7475	0.7356	0.7738
	SPDQ	0.6469	0.6659	0.7172	0.7239	0.6338	0.7531	0.7703	0.7792
	SSDQ	**0.7250**	**0.7631**	**0.8212**	**0.8315**	**0.7067**	**0.8144**	**0.8480**	**0.8613**
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
Image to Text	CVH	0.3942	0.4083	0.4097	0.4154	0.3556	0.3953	0.4087	0.4020
	SCM	0.4768	0.4789	0.4974	0.4919	0.4706	0.4819	0.5088	0.5857
	CMSSH	0.3691	0.3933	0.4097	0.4223	0.3636	0.3955	0.3934	0.4155
	SePH	0.4842	0.4985	0.5095	0.5333	0.5304	0.5483	0.5738	0.5844
	CMCQ	0.5026	0.5429	0.5521	0.5846	0.5434	0.5818	0.5861	0.6265
	DCMH	0.5254	0.5337	0.6061	0.6195	0.5762	0.6278	0.6284	0.6763
	CDQ	0.5536	0.5855	0.6065	0.6273	0.6191	0.6437	0.6727	0.6879
	SPDQ	0.5648	0.5929	0.6245	0.6377	0.6305	0.6748	0.6925	0.6993
	SSDQ	**0.6443**	**0.7092**	**0.7450**	**0.7654**	**0.7529**	**0.7834**	**0.7980**	**0.8045**

Table 1: Mean average precision (MAP@50) comparisons for nine methods on NUS-WIDE (bold font highlights the winner).

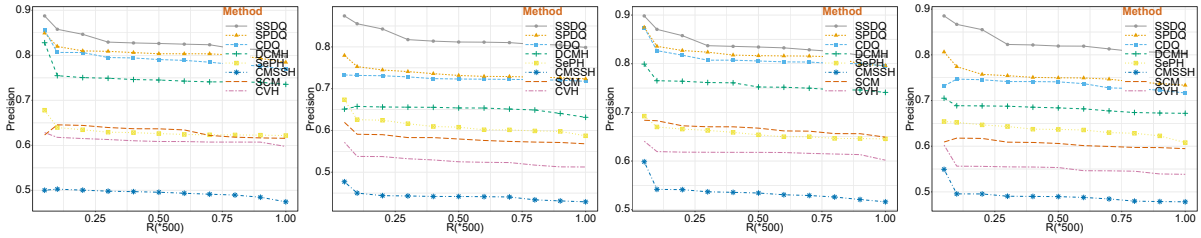
Task	Method	Flickr1M (GIST+SIFT)				Flickr1M (VGG16)			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
Text to Image	CVH	0.6014	0.5892	0.5987	0.6203	0.6275	0.6314	0.6246	0.6248
	SCM	0.6086	0.6160	0.6339	0.6401	0.6465	0.6583	0.6615	0.6789
	CMSSH	0.4796	0.4857	0.4935	0.5047	0.4956	0.5068	0.5111	0.5388
	SePH	0.6059	0.5964	0.6121	0.6436	0.6326	0.6080	0.6359	0.6686
	CMCQ	0.6413	0.6687	0.6914	0.7070	0.7217	0.6849	0.7384	0.7130
	DCMH	0.6048	0.6410	0.7415	0.7620	0.6492	0.6710	0.7683	0.7705
	CDQ	0.6590	0.7332	0.7712	0.8101	0.6852	0.7938	0.8015	0.8237
	SPDQ	0.6665	0.7709	0.7961	0.8219	0.6910	0.8036	0.8111	0.8367
	SSDQ	**0.7740**	**0.8276**	**0.8381**	**0.8588**	**0.7966**	**0.8345**	**0.8480**	**0.8716**
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
Image to Text	CVH	0.4794	0.4918	0.5047	0.5409	0.5140	0.4990	0.5143	0.5641
	SCM	0.5325	0.5787	0.5825	0.5882	0.5953	0.6031	0.6186	0.6211
	CMSSH	0.4284	0.4336	0.4466	0.4517	0.4449	0.4625	0.4575	0.4982
	SePH	0.5806	0.5863	0.6058	0.6201	0.6016	0.6011	0.6211	0.6445
	CMCQ	0.6022	0.6235	0.6256	0.6437	0.6274	0.6328	0.6527	0.6639
	DCMH	0.6252	0.6480	0.6570	0.6663	0.6772	0.6839	0.6919	0.7005
	CDQ	0.6768	0.6851	0.7334	0.7403	0.6920	0.7096	0.7346	0.7486
	SPDQ	0.6962	0.7135	0.7475	0.7537	0.7172	0.7405	0.7627	0.7692
	SSDQ	**0.7557**	**0.8121**	**0.8415**	**0.8500**	**0.7839**	**0.8393**	**0.8477**	**0.8546**

Table 2: Mean average precision (MAP@50) comparisons for nine methods on Flickr1M (bold font highlights the winner).



(a)SIFT, text to image (b)SIFT, image to text (c)VGG, text to image (d)VGG, image to text

Figure 2: Precision with different numbers of top retrieved items on NUS-WIDE with code length 32



(a)SIFT, text to image (b)SIFT, image to text (c)VGG, text to image (d)VGG, image to text

Figure 3: Precision with different numbers of top retrieved items on Flickr1M with code length 32

task of retrieving images given texts with SIFT features, subfigure (b) shows the results for the task of retrieving texts given images with SIFT features, subfigure (c) illustrates the results for the task of retrieving images given texts with VGG16 features and subfigure (d)

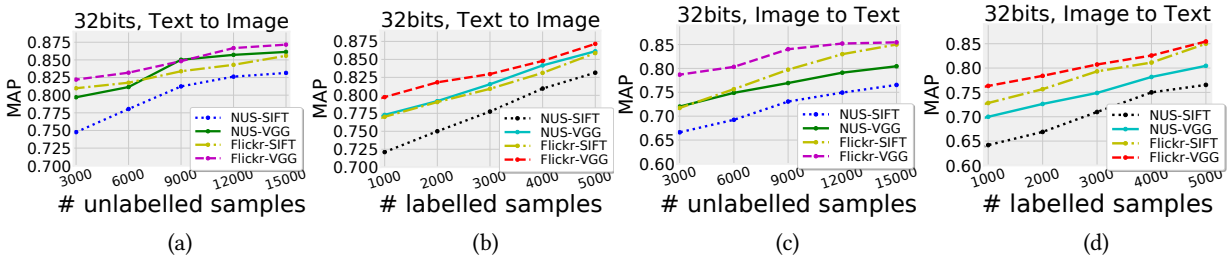


Figure 4: The influence of varying the number of available unlabelled and labelled samples on the performance of SSDQ

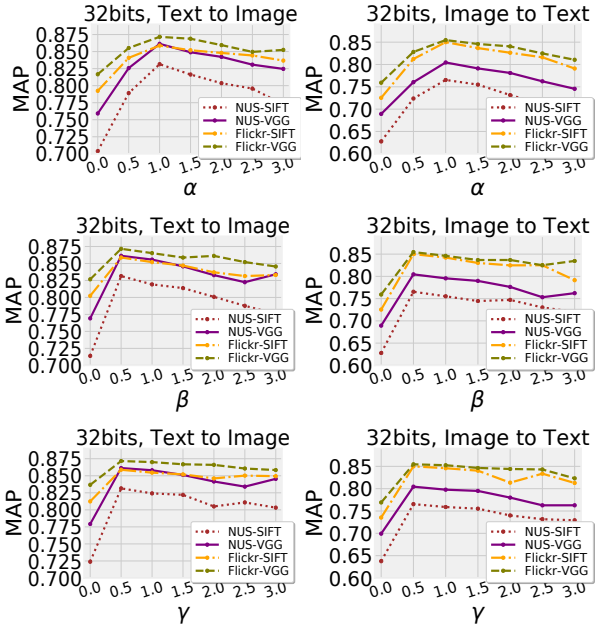


Figure 5: The study on parameter sensitivity for SSDQ

shows the results for the task of retrieving texts given images with VGG16 features. It is obvious that our proposed SSDQ model beats all its comparative partners in all scenarios, which is consistent with the MAP@50 results presented in Table 1 and Table 2.

Varying quantities of unlabelled and labelled samples. As our SSDQ model requires three input sources, i.e., paired data, labelled data and unlabelled data, we evaluate how the number of unlabelled and labelled samples in training procedure affects the search quality in Figure 4. Figure 4(a) and Figure 4(c) display the impact of available unlabelled training samples on the search quality of SSDQ for the task of image retrieval (given texts) and text retrieval (given images) respectively. We can observe that there exists an improvement in the search quality of SSDQ when more and more unlabelled data becomes available for training on all three datasets for both tasks. Similarly, the impact of available labelled training samples on SSDQ's search quality is presented in Figure 4(b) and Figure 4(d), which shows a similar trend in the improvement. A further comparison between Figure 4(a) and Figure 4(b) (also Figure 4(c) and Figure 4(d)) illustrates the improvement in *MAP*@50 brought by the increasing number of labelled samples has a sharper upward slope than that brought by the increasing number of unlabelled samples. This is probably due to the reason that labelled samples can provide extra supervised guidance in determining the similarities between two cross-modal samples. Besides, we would like to point out that the number of training samples is increased simultaneously and equally for both domains (image and text) to avoid information from one modality dominating the other.

Sensitivities of controlling parameters. Last but not least, we test the sensitivity of controlling parameters on all datasets for both tasks in Figure 5. As shown in Equation (8), α, β, and γ control the relative importance of *supervised quantization loss*, *pairwise similarity loss* and *unsupervised distribution match loss* respectively. Figure 5 demonstrates that our proposed SSDQ model is generally insensitive to different parameter settings. Moreover, we observe that the value of optimal α is greater than the value of optimal β and γ, which emphasises the importance of labelled data in cross-modal similarity search. This again validates our motivation for utilizing labelled data to enhance the cross-modal search quality.

6 CONCLUSIONS

Cross-modal similarity search is quite an interesting and practical research topic in both academy and industry. We resort to the combination of compact coding solution and deep structure representation for fast and accurate similarity search in this paper. We present a a semi-supervised deep quantization model (SSDQ) that is capable of simultaneously considering three categories of information, i.e., paired data, labelled data and unlabelled data, for fast and accurate similarity search across different domains (modalities). The proposed model aggregates three losses (each of which is designed to handle one type of information accordingly) together to form an overall cross-modal loss and place it on the top level representation of the neural network for joint optimization. Experiemnts on real-world datasets have demonstrated the advantages of the proposed SSDQ model over existing approaches.

ACKNOWLEDGMENTS

This research is supported by China Postdoctoral Science Foundation No. BX201700136 , National Program on Key Basic Research Project No. 2015CB352300, National Natural Science Foundation of China Major Project No. U1611461 and Shenzhen Nanshan District Ling-Hang Team Grant under No.LHTD20170005. Wenwu Zhu is the corresponding author.

A DATASET PREPROCESSING

NUS-WIDE [6] is a public web image dataset containing 269648 web images annotated by 5018 unique tags in total. We extract images as well as their tags from 10 categories including *bird, building, car, cat, dog, fish, horse, flower, mountain and plane*. For objects in text domain, the most frequent 1000 tags are used to constitute the 1000-dimensional tag occurrence input vectors. For objects in image domain, the 500-dimensional SIFT [19] features and 4096-dimensional VGG16 features [30] pre-trained on ImageNet are adopted as input separately, resulting in two variants of the dataset which are denoted as NUS-WIDE (SIFT) and NUS-WIDE (VGG16) respectively.

Flickr1M [20] contains 1M images associated with tags from Flickr, 25K of which are labelled with 38 concepts. In text domain, we select the most frequent 1000 tags and construct a 1000-dimensional vector extracted from tag occurrences to represent each object. In image domain, we adopt a 3857-dimensional vector consisting of local SIFT and global GIST features [35] as the representation for each object, denoted as Flickr1M (SIFT). Similarly, we also have a Flickr1M (VGG16) variant with 4096-dimensional pre-trained VGG16 features as input.

The statistics of each datasets are shown in Table 3.

Dataset	$\lvert X^{uL}\rvert$	$\lvert X^{L}\rvert$	$\lvert Y^{uL}\rvert$	$\lvert Y^{L}\rvert$	$\lvert X^{P}\rvert(\lvert Y^{P}\rvert)$	#Query
NUS-WIDE	15000	5000	15000	5000	5000	1000
Flickr1M	15000	5000	15000	5000	5000	1000

Table 3: Statistics of NUS-WIDE and Flickr1M

B PARAMETER SETTING

For both NUS-WIDE and Flickr1M, we adopt a 5-layer deep structure for image domain and a 4-layer deep structure for text domain, which is inspired by Wang et al.'work [36]. Detailed settings of the deep model for each dataset are presented in Table 4. We conduct cross-validation on the training data to set $\alpha = 1, \beta = \gamma = 0.5$. Furthermore, the learning rate, the decay and momentum are set to 0.0001, 0.8 and 0.8 respectively.

Dataset	Image Domain	Text Domain
NUS-WIDE (SIFT)	$500-512-128-128-64$	$1K-512-128-64$
NUS-WIDE (VGG16)	$4096-1024-256-128-64$	$1K-512-128-64$
Flickr1M (GIST+SIFT)	$3857-1024-256-128-64$	$1K-512-128-64$
Flickr1M (VGG16)	$4096-1024-256-128-64$	$1K-512-128-64$

Table 4: Deep model settings for NUS-WIDE and Flickr1M

C SUPPLEMENTARY EXPERIMENTAL RESULTS

The visualized results for *Precision v.s. Number of Top Retrieved Items on NUS-WIDE and Flickr1M with code length 16* are shown in Figure 6 and Figure 7.

REFERENCES

[1] Julian Besag. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)* (1986), 259–302.

[2] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3594–3601.

[3] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. 2017. Collective Deep Quantization for Efficient Cross-Modal Retrieval.. In *AAAI*. 3974–3980.

[4] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1445–1454.

[5] Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu. 2016. Correlation autoencoder hashing for supervised cross-modal search. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 197–204.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*. ACM, 48.

[7] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2075–2082.

[8] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. 2013. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2916–2929.

[9] Yao Hu, Zhongming Jin, Hongyi Ren, Deng Cai, and Xiaofei He. 2014. Iterative multi-view hashing for cross media indexing. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 527–536.

[10] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2011. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2011), 117–128.

[11] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE.

[12] Saehoon Kim, Yoonseop Kang, and Seungjin Choi. 2012. Sequential spectral learning to hash with multiple representations. In *European Conference on Computer Vision*. Springer, 538–551.

[13] Shaishav Kumar and Raghavendra Udupa. 2011. Learning hash functions for cross-view similarity search. In *IJCAI proceedings-international joint conference on artificial intelligence*, Vol. 22. 1360.

[14] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3864–3872.

[15] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. 2012. Supervised hashing with kernels. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2074–2081.

[16] Xianglong Liu, Junfeng He, Cheng Deng, and Bo Lang. 2014. Collaborative hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2139–2146.

[17] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S Yu. 2016. Composite correlation quantization for efficient multimodal retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 579–588.

[18] Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip. 2014. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 26, 5 (2014), 1076–1089.

[19] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. IEEE, 1150–1157.

[20] B. Thomee Mark J. Huiskes and Michael S. Lew. 2010. New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*. ACM, New York, NY, USA, 527–536.

[21] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber. 2014. Multimodal similarity-preserving hashing. *IEEE transactions on pattern analysis and machine intelligence* 36, 4 (2014), 824–830.

[22] Sean Moran and Victor Lavrenko. 2015. Regularised cross-modal hashing. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 907–910.

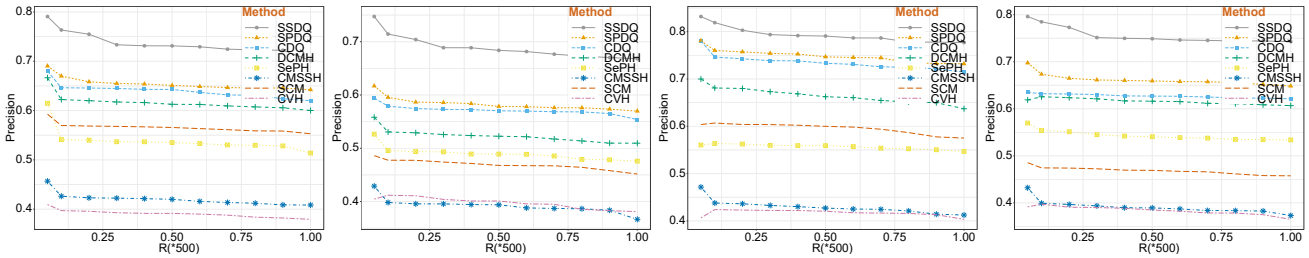
[23] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[24] Mohammad Norouzi and David M Blei. 2011. Minimal loss hashing for compact binary codes. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 353–360.

[25] Mohammad Norouzi and David J Fleet. 2013. Cartesian k-means. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 3017–3024.

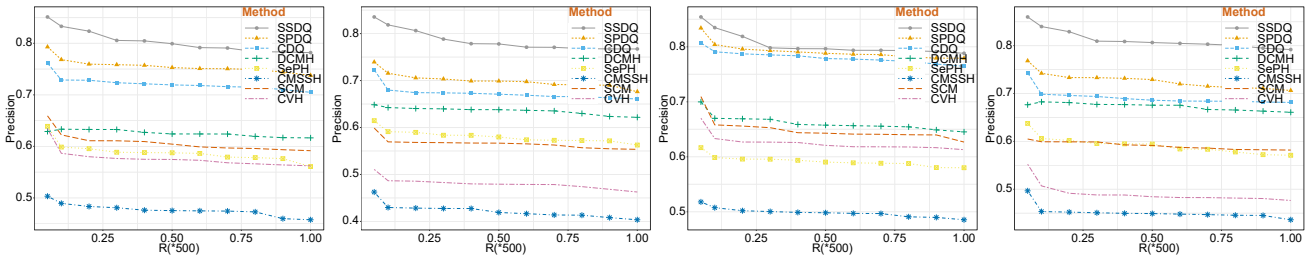
[26] Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Comparing apples to oranges: a scalable solution with heterogeneous hashing. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 230–238.

[27] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50, 7 (2009), 969–978.



(a)SIFT, text to image (b)SIFT, image to text (c)VGG, text to image (d)VGG, image to text

Figure 6: Precision with different numbers of top retrieved items on NUS-WIDE with code length 16



(a)SIFT, text to image (b)SIFT, image to text (c)VGG, text to image (d)VGG, image to text

Figure 7: Precision with different numbers of top retrieved items on Flickr1M with code length 16

[28] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* (2013), 2263–2291.

[29] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. 2015. Supervised Discrete Hashing.. In *CVPR*, Vol. 2. 5.

[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[31] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. 2013. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia* 15, 8 (2013), 1997–2008.

[32] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 785–796.

[33] David Sontag, Amir Globerson, and Tommi Jaakkola. 2011. Introduction to dual composition for inference. In *Optimization for Machine Learning*. MIT Press.

[34] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, Vol. 79.

[35] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[36] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. 2015. Deep Multimodal Hashing with Orthogonal Regularization.. In *IJCAI*, Vol. 367. 2291–2297.

[37] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. 2012. Semi-supervised hashing for large-scale search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 12 (2012), 2393–2406.

[38] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. 2014. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927* (2014).

[39] Xiaojuan Wang, Ting Zhang, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang. 2016. Supervised quantization for similarity search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018–2026.

[40] Yair Weiss, Antonio Torralba, and Rob Fergus. 2009. Spectral hashing. In *Advances in neural information processing systems*. 1753–1760.

[41] Botong Wu, Qiang Yang, Wei-Shi Zheng, Yizhou Wang, and Jingdong Wang. 2015. Quantized Correlation Hashing for Fast Cross-Modal Search.. In *IJCAI*. 3946–3952.

[42] Fei Wu, Zhou Yu, Yi Yang, Siliang Tang, Yin Zhang, and Yueting Zhuang. 2014. Sparse multi-modal hashing. *IEEE Transactions on Multimedia* 16, 2 (2014), 427–439.

[43] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. 2014. Supervised hashing for image retrieval via image representation learning.. In *AAAI*, Vol. 1. 2.

[44] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. 2018. Shared Predictive Cross-Modal Deep Quantization. *IEEE Transactions on Neural Networks and Learning Systems* 99 (2018), 1–12.

[45] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 395–404.

[46] Dongqing Zhang and Wu-Jun Li. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization.. In *AAAI*, Vol. 1. 7.

[47] Dan Zhang, Fei Wang, and Luo Si. 2011. Composite hashing with multiple information sources. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 225–234.

[48] Ting Zhang, Chao Du, and Jingdong Wang. 2014. Composite Quantization for Approximate Nearest Neighbor Search.. In *ICML*. 838–846.

[49] Ting Zhang and Jingdong Wang. 2016. Collaborative quantization for cross-modal similarity search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2036–2045.

[50] Yi Zhen and Dit-Yan Yeung. 2012. Co-regularized hashing for multimodal data. In *Advances in neural information processing systems*. 1376–1384.

[51] Yi Zhen and Dit-Yan Yeung. 2012. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 940–948.

[52] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 415–424.

[53] Xiaojin Zhu. 2005. Semi-supervised learning literature survey. (2005).

[54] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 143–152.