SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2017, Vol. 60 092104:1–092104:14
doi: 10.1007/s11432-015-0902-2

A novel cross-modal hashing algorithm based on multimodal deep learning

Wen QU[1], Daling WANG[1,2], Shi FENG[1,2], Yifei ZHANG[1,2] & Ge YU[1,2*]

[1]*School of Information Science and Engineering, Northeastern University, Shenyang* 110819*, China;*
[2]*Key Laboratory of Medical Image Computing, Shenyang* 110819*, China*

Received April 29, 2016; accepted August 15, 2016; published online March 21, 2017

Abstract With the growing popularity of multimodal data on the Web, cross-modal retrieval on large-scale multimedia databases has become an important research topic. Cross-modal retrieval methods based on hashing assume that there is a latent space shared by multimodal features. To model the relationship among heterogeneous data, most existing methods embed the data into a joint abstraction space by linear projections. However, these approaches are sensitive to noise in the data and are unable to make use of unlabeled data and multimodal data with missing values in real-world applications. To address these challenges, we proposed a novel multimodal deep-learning-based hash (MDLH) algorithm. In particular, MDLH uses a deep neural network to encode heterogeneous features into a compact common representation and learns the hash functions based on the common representation. The parameters of the whole model are fine-tuned in a supervised training stage. Experiments on two standard datasets show that the method achieves more effective results than other methods in cross-modal retrieval.

Keywords hashing, cross-modal retrieval, cross-modal hashing, multimodal data analysis, deep learning

Citation Qu W, Wang D L, Feng S, et al. A novel cross-modal hashing algorithm based on multimodal deep learning. Sci China Inf Sci, 2017, 60(9): 092104, doi: 10.1007/s11432-015-0902-2

1 Introduction

The growing popularity of social media on Web 2.0 in recent years has led to a dramatic increase in the amount of multimodal data. For example, photographs are usually associated with captions and tags, videos contain visual and audio signals, and tweets often consist of text, images, and videos. At the same time, when users search the Internet and acquire information, they want to obtain a comprehensive result consisting of multiple types of media. Traditional information retrieval systems use text alone as query input, and so image and video retrieval in most such systems is based on text queries. With the rapid development of mobile equipment such as telephones and tablet computers, users may now perform queries using image, audio, or video input rather than text [1, 2]. Therefore, there is an emerging need for retrieval and search methods based on data entities from multiple modalities.

As a method allowing a system to handle large amounts of multimedia data, hashing has attracted increasing attention owing to its advantages in reducing both computational cost and storage requirements. Most existing hashing methods are designed for unimodal data, such as image hashing [3] and

* Corresponding author (email: yuge@ise.neu.edu.cn)

In Shijiazhuang, a police car collided with a bus, which cause one policeman dead and two injured.

The life without plan is the failure of the life!
The life without goal is a sad life!
The life without ideal is a confused life!
The life without friendship is a lonely life!
The life without love is a lost life!

Figure 1 (Color online) Two examples of multimodal data downloaded from Tencent Weibo. (a) The text and image of a tweet describe the same concept; (b) the two modalities of a tweet contain different semantics.

video hashing [4, 5]. The most well known of these is locality-sensitive hashing (LSH) [6], which uses random projections to obtain the hash functions. However, LSH usually needs quite a long hash code and hundreds of hash tables to guarantee good retrieval performance. To address these problems, several data-dependent learning based methods have been proposed [7]. In recent years, the growth in real-world applications has led to cross-media retrieval becoming an important research topic. Much work has been devoted to extending unimodal hashing to multimodal settings [8]. Cross-modal hashing maps data from different modalities into a shared Hamming space in which the distance between similar objects is small. In this Hamming space, all data are represented as hash codes and can be searched quickly even for databases with millions of data. Compared with unimodal hashing, multimodal hashing preserves both intra- and inter-modal similarity in the Hamming space. The cross-modal hash function preserves not only the respective information in each modality but also the mutual information in different modalities.

Most previous cross-modal hashing methods have been based on the assumption that the multimodal data used for hash function learning are available in all the modalities and are semantically consistent across different modalities. Therefore, these methods are unable to make use of unlabeled data or multimodal data with missing values. In real-world applications, however, data on the Internet are very noisy and may have missing modalities. Figure 1 shows two examples of multimodal data downloaded from Tencent Weibo. Figure 1(a) shows an ideal situation with regard to multimodal data, in which the text and image describe the same concept. In contrast, as shown in Figure 1(b), the two modalities of a tweet contain different semantics. Furthermore, the data generated by users usually have values missing from some modality (e.g., some pictures uploaded by users lack any tags or words). On the other hand, many approaches represent multimodal data through clustering [9] or dictionary learning [10], which construct the corresponding alignments for pair matching between modalities. When a new modality is added to the system, its relationship with each existing modality has to be learned again.

To address these problems, in this paper, we propose a multimodal deep-learning-based hashing (MDLH) algorithm, which learns the common feature space of different modalities using a deep neural network. Multimodal deep learning can learn a compact and robust "semantic" representation of multimodal data, and is able to handle both noisy data and data with missing modalities. Experiments on two realistic datasets show that the proposed method can realize cross-modal hashing effectively.

The rest of the paper is organized as follows: In Section 2, we review related work. In Section 3, we elaborate the proposed method. In Section 4, we demonstrate the use of our approach in cross-modal retrieval and describe the experimental results. Finally, we conclude the paper in Section 5.

2 Related work

Previous work on cross-modal hashing and multimodal deep learning will be reviewed in the following subsections.

2.1 Cross-modal hashing

Hashing can be categorized as unimodal hashing, multimodal hashing, or cross-modal hashing. In unimodal hashing, the most well-known methods are LSH [6] and spectral hashing [7]. Multimodal hashing compares the multimodal features of data and returns the search results for each modality. For example, when retrieving an image according to multimodal descriptors (color, scale-invariant feature transform (SIFT), and bag of words (BOW)), multimodal hashing projects each feature into a Hamming space and combines the multiple results together. Cross-modal hashing focuses on analyzing the relationship between modalities and provides cross-modal queries. For example, given the color feature of an image as the input, the system returns results according to the SIFT descriptor. Here, the modality is the feature or media type, so cross-modal in this context means cross-feature or cross-media.

Existing unimodal data hashing methods involve two steps. First, the original data are projected into a low-dimensional space. Then, the new representation is quantized into hash codes. For unsupervised situations, many embedding methods have been proposed, such as random projection [6] and spectral decomposition [7]. Multimodal data hashing also involves two steps, but with more restrictions. Bronstein et al. [11] proposed the first cross-modal hashing model, cross-modality similarity-sensitive hashing (CMSSH). Given two modalities, CMSSH learns two groups of hash functions that are such that similar data (in different modalities) are separated by smaller distances in the Hamming space, while dissimilar data (in different modalities) are separated by larger distances. CMSSH retains the relationship between data in different modalities, but ignores similarities in the same modality. Kumar and Udupa [12] extended spectral hashing to a multimodal setting and proposed cross-view hashing (CVH), which minimizes the distance between similar data both in the same modality and in different modalities. In multimodal latent binary embedding (MLBE) [8], a probability-generating model is used to represent the data, and the latent factors learned are used as the hash codes. There is no independent restriction on the hash codes, so these may have a high redundancy. In the approach adopted by Yu et al. [10], dictionary learning is used to represent data in different modalities, and the hash function is learned based on sparse codes. Dictionaries for different modalities are connected through a coupled dictionary space. In iterative multiview hashing (IMVH) [13], both intra- and inter-similarity of the data are retained. Song et al. [14] proposed inter-media hashing, in which a set of corresponding images and text are used as the inter-media to learn the relations among multiple modalities. Unlike most approaches, in which optimization of the quantizer is performed independently of correlation modeling, quantized correlation hashing (QCH) [15] optimizes the two processes simultaneously. Owing to their ability to capture high-level representations, deep learning techniques have shown advantages in describing multimodal data. Kang et al. [16] exploited deep networks for hashing and proposed deep multiview hashing (DMVH). Wang et al. [17] imposed an orthogonal regularizer on the weighting matrices of the model to reduce the amount of redundant information lying in the multimodal representations.

These methods learn cross-modal relations with different techniques. However, most existing multimodal hashing algorithms assume that each data example appears in all modalities. Wang et al. [18] proposed a hashing approach to deal with partial multimodal data, in which data consistency among different modalities is ensured via latent subspace learning and inter-similarity is preserved through the use of a graph Laplacian.

2.2 Multimodal deep learning

In deep learning, a layer network structure is constructed to simulate the human brain, and representations for data are learned from the bottom upward. Each layer of the network corresponds to a representation. Recently, deep learning has been widely used in many applications, including speech recognition [19], face

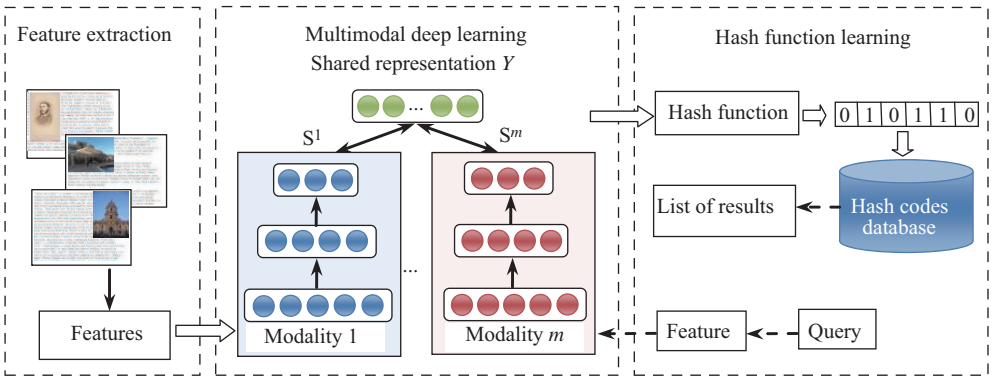

Figure 2 (Color online) Framework of the multimodal feature deep learning hashing.

recognition, image classification [20], and object recognition, and impressive results have been achieved. Representative deep learning approaches include deep belief networks (DBN), autoencoders, stacked denoising autoencoders, deep Boltzmann machines (DBMs), and deep energy models. Ngiam et al. [21] used a DBM to learn cross-modal representations of video and audio data, and reconstructed the data of a missing modality. Srivastava and Salakhutdinov [22] proposed a DBN to learn representations of multimodal data. Sohn et al. [23] proposed an improved multimodal deep learning model.

This previous work has focused on solving the data reconstruction problem in the case of a missing modality. Our work focuses on learning the relations between different modalities and proposing a semantic and common representation of multimodal data. The work most similar to ours is that by Wu et al. [24], in which deep learning is used to learn the optimal combination of different modalities. However, in contrast to their approach, we focus on learning a common representation of multimodal data using a deep neural network.

Deep learning techniques have also been used to learn low-dimensional representations for multimodal retrieval. Wang et al. [25] proposed an effective mapping technique in which multimodal stacked autoencoders (MSAEs) are used to project high-dimensional features into a common low-dimensional space. To solve the problem of cross-modal retrieval, a correspondence autoencoder (corr-AE) [26] has been proposed for correlating hidden representations of two unimodal autoencoders.

3 MDLH algorithm

In this section, we present the details of the MDLH algorithm. Figure 2 shows the framework of our method. First, the multimodal features of multimodal data are extracted as inputs. Then, we use the multimodal deep learning method to learn the common representation for them. Finally, the hash function of each modality is used to map the data into the Hamming space. In the following, the notation and problem formulation are introduced, and the model of multimodal deep learning is then presented, followed by a description of the hashing function learning.

3.1 Notation and problem definition

We are given a set of multimodal data $O = \{O^1, \ldots, O^p, \ldots, O^M\}$ $(p = 1, \ldots, M)$ consisting of N multimodal data, where O_i is the ith datum in O. Each multimodal data $O_i = \{O_i^1, \ldots, O_i^p, \ldots, O_i^M\}$ contains M modalities, where O_i^p $(p = 1, \ldots, M)$ is the pth modality of O_i. We extract different features for different modalities, and use $X^p = (x_1^p, \ldots, x_{N_p}^p) \in \Re^{D_p \times N_p}$ to represent the features of the pth modality, where D_p is the dimension of the feature space and N_p is the number of data in the pth modality. MDLH aims to learn a series of hash functions $X^p \to B^p = (0, 1)^c$ for which similar data have similar hash codes both intra-and inter-modally. Here, c is the length of the hash code. Note that all the modalities have the same length.

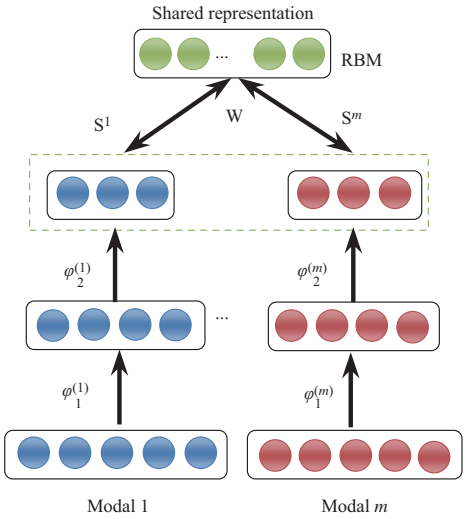

Figure 3 (Color online) Multimodal deep learning model.

As in the middle part of Figure 2, we project the original multimodal features to the shared space via multimodal deep learning. We denote the shared representation of the multimodal data by S and define the projections as

$$f_1^p : X^p \to S^p, \qquad f_2^p : S^p \to Y^p, \tag{1}$$

where f_1^p and f_2^p denote the projections for each modality and cross-modality, respectively. Data for each individual modality are first converted into representations for a single modality, denoted by S^p. The process aims to preserve the intra-similarity for each modality. Then, data of all modalities, represented by S^p, are mapped into a common space Y^p, where the relations between multiple modalities are learned. After the shared representations have been learned, they are mapped into Hamming space using a linear projection:

$$g^p : Y^p \to B^p. \tag{2}$$

We learn the projection such that inter-similarity is preserved and then transform the values in binary form into Hamming space. We take a training dataset $T = (x_i^{m_i}, x_j^{m_j})^k$ $(k = 1, \ldots, K)$ from O, where $x_i^{m_i}$ and $x_j^{m_j}$ are the features of $o_i^{m_i} \in O^{m_i}$ and $o_i^{m_j} \in O^{m_j}$, respectively. L_{ij} is an indicator that is equal to 1 if two data $o_j^{m_i}$ and $o_j^{m_j}$ belong to the same class; otherwise $L_{\mathrm{ij}} = -1$. The distance between two data in Hamming space is defined as

$$d(x_i^{m_i}, x_j^{m_j}) = \|B_i^{m_i} - B_j^{m_j}\|_F^2. \tag{3}$$

We then formulate the problem of learning the projection as the following optimal problem with object function:

$$\min_f \sum_{k=1}^{K} L_{\mathrm{ij}} d\left(x_i^{m_i}, x_j^{m_j}\right). \tag{4}$$

3.2 Feature learning based on multimodal deep learning

Multimodal deep learning consists of two components: (1) feature learning for each single modality; (2) shared feature learning for multimodal features. In the following, we first describe the multimodal deep learning model and compare it with other models. Then we describe the two components of the model.

3.2.1 *Multimodal deep learning models*

Figure 3 shows the deep neural network for multimodal deep learning. The whole model is learned in three steps: First, the unlabeled data U of each modality is used to pre-train the deep learning network

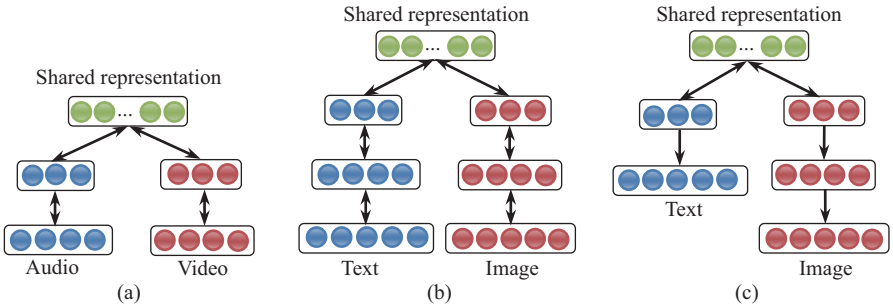

Figure 4 (Color online) Multimodal deep learning models based on RBMs. (a) Deep RBM; (b) multimodal DBM; (c) multimodal DBN.

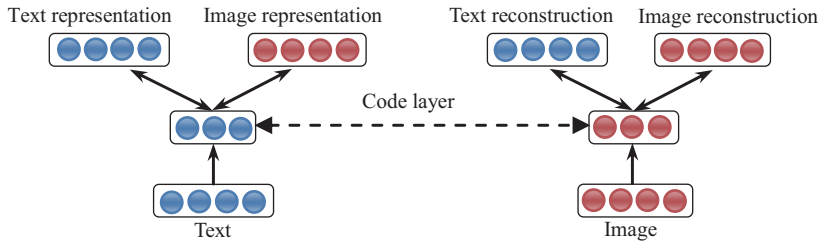

Figure 5 (Color online) Correspondence cross-modal autoencoder.

using a stacked denoising autoencoder (SDA) (see Subsection 3.2.2). Then, the multimodal data O is represented using the SDA of each modality, and the outputs are input into the restricted Boltzmann machine (RBM) to learn the relationship between multiple modalities. Finally, the training dataset T is used to update the parameters of the model.

There are several multimodal deep learning models, including deep autoencoders [21], multimodal DBMs [22], multimodal DBNs (mDBNs) [17] and correspondence cross-modal autoencoders (corr-cross-AEs) [26]. Ngiam et al. [21] used a deep RBM (Figure 4(a)) to learn features over multiple modalities. They trained the deep autoencoder network using an augmented dataset that had only a single modality as input. Their work was aimed at learning high-dimensional latent features to perform discriminative classification tasks. Srivastava and Salakhutdinov [22] extended an mDBM to model joint distributions over image and text inputs. They used separate two-layer DBMs for each modality, as shown in Figure 4(b), and combined the two models by adding an additional layer of binary hidden units on top of them. Another way of using a deep model to combine multimodal inputs is through an mDBN [17] (Figure 4(c)) that is composed of a DBN for each modality, with the joint RBM capturing the correlations among multiple modalities. The corr-cross-AE [26] replaces the basic autoencoders by cross-modal autoencoders, as illustrated in Figure 5. All of these models use only one kind of basic unit, whereas our model combines the autoencoder with the RBM. The deep autoencoder and corr-cross-AE use only autoencoders, while the DBM and mDBN use only use RBMs in the network structure. The proposed multimodal deep learning framework adopts autoencoders for individual modalities, and it models the multimodal relation with an RBM.

3.2.2 *Feature learning for a single modality*

For different modalities, data have different feature representations and statistical properties in the low-level feature space. For example, images can be described with color, SIFT visual features, and so on, while the text surrounding the images can be represented using the the BOW feature. These low-level features are diverse in dimension and representation (e.g., SIFT with 128 dimensions and color with 512 dimensions) and have a "semantic gap" with the high-level semantic concepts. Therefore, we use a modality-specific structure to learn features for each modality separately, with the aim of learning a compact and robust high-level representation. The higher level of the structure corresponds to high-level

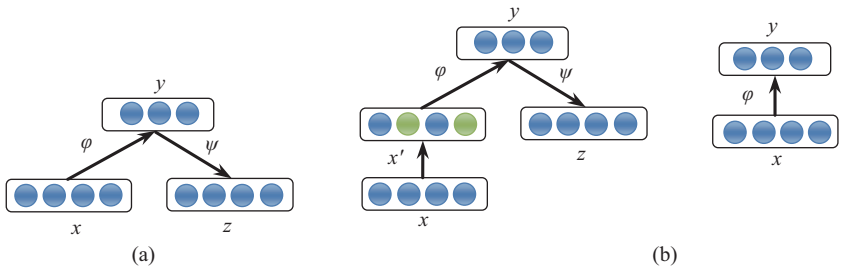

Figure 6 (Color online) Autoencoder and denoising autoencoder. (a) Autoencoder; (b) denoising autoencoder.

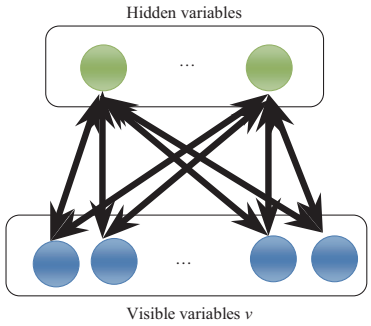

Figure 7 (Color online) RBM.

features, which are further correlated for multimodal feature learning.

We use a SDA [27] as the modality-specific structure, which is a type of autoencoder. Autoencoders have been widely used in unsupervised feature learning and classification tasks. They can be seen as special neural networks with three layers: an input layer, a latent layer, and a reconstruction layer (Figure 6(a)). A denoising autoencoder (DAE) adds noise to the training data, as shown in Figure 6(b). First, a noisy version of x is constructed through a stochastic mapping. Then the noisy version x' is mapped through an autoencoder to a hidden representation $y = \varphi(x')$, and y is used to reconstruct a clean version of x by $z = \psi(y)$. We use a nonlinear one-layer neural network as the unit of the SDA, with the encoding function being given by

$$y = \varphi(x) = \text{sigmoid}(Qx + r) \tag{5}$$

and the decoding function by

$$z = \psi(y) = \text{sigmoid}(Q'y + t). \tag{6}$$

Several DAEs are stacked to build a layer structure in which the output of each layer is the input to the layer above. Once the encoding function has been learned, the encoding function is no longer needed.

3.2.3 *Multimodal feature learning*

After learning the representation of each modality, we use an RBM [28] to model the relations between different modalities and learn their shared representation.

An RBM is an undirected graphical modal with stochastic visible unit v and stochastic hidden unit h (Figure 7). Each visible unit connects to each hidden unit, but there are no connections among hidden variables or among visible variables. The structure of the model is shown in Figure 6. The model defines the following energy function E:

$$E(v, h; \theta) = -a^{\mathrm{T}}v - b^{\mathrm{T}}h - v^{\mathrm{T}}Wh, \tag{7}$$

where $\theta = a$, b, and W are the model parameters. The joint distribution over the visible and hidden units is defined by

$$p(v, h; \theta) = \frac{1}{Z(\theta)} \exp[-E(v, h; \theta)], \tag{8}$$

where $Z(\theta)$ is a normalization constant. The jth hidden node is set equal to 1 with probability

$$p(h_j|v) = \text{sigmoid}\left[\frac{1}{\sigma^2}(b_j + W_j^{\mathrm{T}} v)\right]. \tag{9}$$

According to the activation probabilities of the hidden units, the model can reconstruct the original data. We minimize the loss function between the original and reconstructed data, and learn the parameter using contrastive divergence [29].

The training procedure for the MDLH is similar to those of other deep networks [30]. We pre-train each layer in a greedy pairwise approach [1]. First, we train the parameters for the bottom two layers and fix the learned parameters. Then we repeat the process for the next two layers, and so on until we reach the top layer.

After pre-training each layer of the MDLH, we fine-tune the model using annotated data, in which each pair of samples describes the same object in two modalities (text and image). In practice, we first clamp examples that have two modalities available, and require the network to reconstruct both modalities. Then we zero-out one of the input modalities separately, still requiring the network to reconstruct both modalities.

The model generates the shared representation by estimating $p(h|v)$. The activation probabilities of hidden units constitute the joint representation of the inputs. After obtaining the shared representation y by means of the multimodal deep learning model, we update the parameters of the last layer. Finally, we use back-propagation [17, 31] to update the parameters in the lower layers of the network.

3.3 Hash function learning

Denoting the shared representation for the data by s, the linear transformation from s to the hash code is

$$g(s) = \text{sign}\left(\boldsymbol{P}^{\mathrm{T}} s\right), \tag{10}$$

where $\boldsymbol{P}$ is the projection matrix.

Denoting by $Y^i = [Y_1^i, \ldots, Y_K^i]$ and $Y^j = [Y_1^j, \ldots, Y_K^j]$ the representations for all datasets, the corresponding hash codes are B_i and B_j, respectively. To preserve the maximal consistency of different modalities, the objective function is defined as

$$\min_{B_i, B_j} \|B_i - B_j\|_F^2. \tag{11}$$

The optimization problem in (11) is equivalent to a balanced graph partitioning problem and is NP-hard. We solve the derived objective function on Y^i and Y^j as

$$\min_{\boldsymbol{P}} \|\boldsymbol{P} * Y^i - \boldsymbol{P} * Y^j\|_F^2. \tag{12}$$

This can be converted into an eigenvalue problem. However, it has a high time complexity and cannot be used to learn hash functions when K is large. To handle this problem, an approach based on sparse characteristics is proposed in [22]. Since our representations Y^i and Y^j are sparse, we follow the method of Ref. [22] to learn the projection matrix $\boldsymbol{P}$. Algorithm 1 summarizes the multimodal deep-learning-based cross-modal hashing. Given new data, the hash code is generated in two steps: First, the features of the data are extracted and multimodal deep learning is used to represent the data. Then, the linear projection function g is used to compute the hash code of the data.

3.4 Extension of new modality

Since different datasets contain different modalities, it is essential that the hashing method be easily extended to more modalities. The proposed MDLH can add new datasets and modalities based on the existing model.

We denote the new multimodal data by $A = \{A_1, \ldots, A_N\}$, which consists of T modalities. In particular, the modality t is a new modality. The procedure of adding the dataset A for training is as follows:

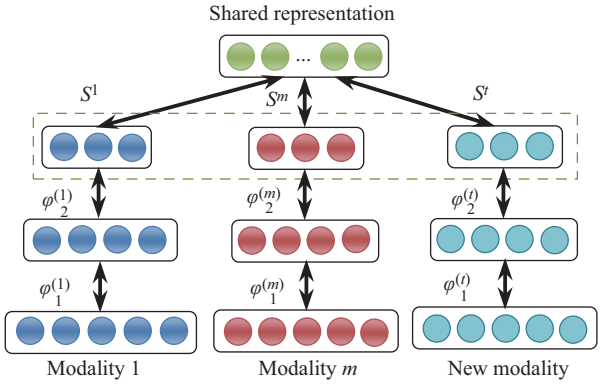

Figure 8 (Color online) Adding a new modality to the existing model.

Algorithm 1 Multimodal deep-learning-based cross-modal hashing

Input: multimodal data O, training data U;
Output: projection f^p, projection g^p;

```
for m = 1 : M do
    pre-training the SDA for modality m;
end for
pre-training the RBM;
while object function is not converged do
    for k = 1 : K do
        (x_i, x_j) ⇐ T;
        update the parameters W, b;
        update the parameter in the lower layer using back-propagation;
    end for
end while
Compute P.
```

(1) Pre-train the stacked autoencoder for the new modality t.

(2) Add the units in the highest layer of the stacked autoencoder to the visible layer of the existing model (as shown in Figure 8).

(3) Fine-tune the MDLH model using data in A with all the modalities.

(4) Learn the hash function as described in Subsection 3.3.

4 Experiments

We evaluate our method on two real-world datasets for a cross-modal similarity search and analyze the results. In detail, the datasets consist of text and images, and we use text as query to search similar images and image as query to search similar texts. First, we introduce the dataset and the experimental setup. We then show the results and compare them with the results of other methods.

4.1 Datasets and settings

Two datasets are used in our experiment: Wikipedia–Picture of the Day and NUS-WIDE. Both include two modalities (pictures and text). Wikipedia–Picture of the Day [32] includes 2866 multimedia documents collected from the Wikipedia website, in which each document includes one picture and at least 70 words. The dataset provides the topic probability of each text on 10 categories, computed using latent Dirichlet allocation (LDA) [33]. Previous experiments used the topic probability as a text feature, but this is too sparse to be a suitable input to deep learning. Therefore, we extract the vector space modality of each text as a feature. The feature of images uses a SIFT descriptor [34] based on a bag-of-visual-words model, which quantizes the descriptors into 1000-dimensional vectors.

The NUS-WIDE dataset is a real-world image dataset collected by the laboratory for media search at the National University of Singapore [35]. It includes 81 categories and 269648 images. Each image

Table 1 Evaluation of the mean average precision of MDLH with different numbers of hidden layers

Number of hidden layers	Wikipedia–Picture of the Day	NUS-WIDE
1-layer	0.3745	0.4398
2-layer	0.3776	0.4502
3-layer	**0.3844**	0.4512
4-layer	0.3783	**0.4565**

Table 2 Number of units in the model

Dataset	Image pathway	Text pathway
Wikipedia–Picture of the Day	1000–512–128	1000–512–128
NUS-WIDE	1000–512–128	1000–512–128

corresponds to multiple tags, and each image–text pair is annotated by at least one category. The image is represented by 1000-dimensional bag-of-visual-words/SIFT descriptors, and the text corresponding to the image is represented by a 1000-dimensional vector of tags.

To select the number of layers and the dimensionality settings of the hidden layers, we evaluate the impact of the number of layers on the deep networks for the proposed MDLH algorithm. Table 1 shows the mean average precision (mAP) performance on two datasets. It can be seen that three-layer deep networks tend to achieve better performance than those with other numbers of layers. For both datasets, the model consists of a three-layer image pathway, a three-layer text pathway, and a joint RBM. In detail, the number of units in each layer is summarized in Table 2.

4.2 Evaluation metrics

We use mAP [8] as the evaluation metric for effectiveness in our experiment. This evaluation metric has been widely used [8,36]. It evaluates the performance of similarity searches, with larger values indicating better performance and with similar results having high ranks. Given a query and R retrieved instances, the average precision (AP) is defined as

$$\mathrm{AP} = \frac{1}{L}\sum_{r=1}^{R} P(r)\delta(r), \tag{13}$$

where L is the number of relevant instances in the result. $P(r)$ is the accuracy of the top r instances. $\delta(r)$ is the indicator function, which is equal to 1 if the rth instance is relevant to the query or 0 otherwise. mAP is the mean of all the AP values, and we set $R = 100$ in our experiments.

4.3 Comparison methods

We compare our method with four other cross-modal hash methods: CMSSH, CVH, latent semantic sparse hashing (LSSH), and IMVH.

• CMSSH [11] constructs two groups of linear hash functions to retain the similarity relationship between different modalities.

• CVH [12] extends unimodal spectral hashing to a multimodal context, retaining the similarity relationship between different modalities and in the same modality.

• LSSH [36] uses matrix factorization and sparse coding to map text and image into the latent factor space separately.

• IMVH [13] retains interior and exterior similarities, while making a clear distinction between data belonging to different categories.

CMSSH and CVH generate different hash codes for different modalities, but they ensure that hash codes in the same modality have the same length.

Table 3 Mean average precision of different methods on the Wikipedia–Picture of the Day dataset

Task	Method	HCL[a)] = 16	HCL = 32	HCL = 64
Image query text	CMSSH	0.3183	0.3275	0.2750
	CVH	0.3140	0.3345	0.2760
	LSSH	0.3730	0.3940	0.3887
	IMVH	0.3812	0.3921	0.3879
	MDLH	**0.3919**	**0.3940**	**0.4030**
Text query image	CMSSH	0.3321	0.3173	0.3147
	CVH	0.3005	0.3322	0.3107
	LSSH	0.3552	0.3559	0.3545
	IMVH	0.3642	0.3624	0.3644
	MDLH	**0.3840**	**0.3729**	**0.3604**

a) HCL: hash code length.

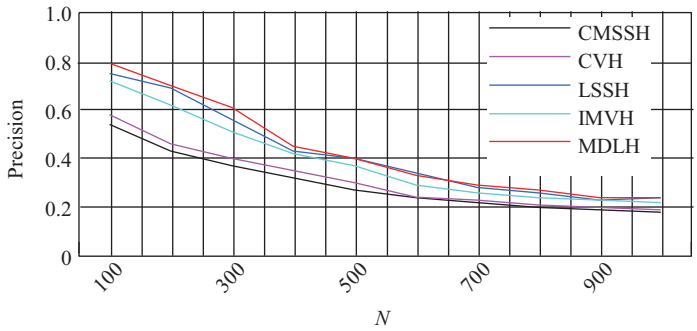

Figure 9 Top-N precision of different methods on the Wikipedia–Picture of the Day dataset.

4.4 Results

4.4.1 *Results on the Wikipedia–Picture of the Day dataset*

We select 90% of the dataset as training data, 5% as unlabeled data, and the rest as the query set for MDLH. Other methods use 95% of the dataset (training data and unlabeled data for MDLH) as training data and the rest as the query set. The mAP values of our method and the other methods are shown in Table 3. We can see that MDLH outperforms most of the other methods for most code lengths. As the code length increases to 64, the performance no longer increases. This phenomenon has also been observed and analyzed elsewhere [8,36]. Since most baseline methods use eigenvalue decomposition and have orthogonality constraints, each bit shows no correlation with the others. Therefore, the first few projection directions may have higher variance than the other projections. As the code length increases, the hash codes will become dominated by bits with low variance. Previous studies reported better performance on the task "text query image" than the task "image query text", because they used topics rather than words as the text feature so that the text queries were represented as the 10 topics, which simplified the research problem. Furthermore, we report the top-N precision curve of the results on the Wikipedia–Picture of the Day dataset in Figure 9, which shows the variation of precision as the number of retrieved instances changes.

4.4.2 *Results on the NUS-WIDE dataset*

Some categories in the NUS-WIDE dataset are scarce, so we select eight categories that contain more instances than the others. We select 90% of the dataset as the training data, 5% as unlabeled data, and 5% as the query data for MDLH. The mAP values of all the methods on NUS-WIDE are shown in Table 4. The performance of all methods increases to some degree on the NUS-WIDE dataset. The reason is that the text modality of the NUS-WIDE dataset consists of tags rather than paragraphs as in the Wikipedia–Picture of the Day dataset. First, the tags have less noise than the paragraphs, which

Table 4 Mean average precision of different methods on the NUS-WIDE dataset

Task	Method	HCL = 16	HCL = 32	HCL = 64
Image query text	CMSSH	0.4405	0.4389	0.3934
	CVH	0.3756	0.3729	0.3619
	LSSH	0.4517	0.4437	0.4460
	IMVH	0.4520	0.4489	0.4446
	MDLH	**0.4526**	**0.4537**	**0.4555**
Text query image	CMSSH	0.4113	0.3984	0.3722
	CVH	0.3805	0.3629	0.3899
	LSSH	0.4271	0.4178	0.4143
	IMVH	0.4189	0.4250	0.4130
	MDLH	**0.4496**	**0.4478**	**0.4485**

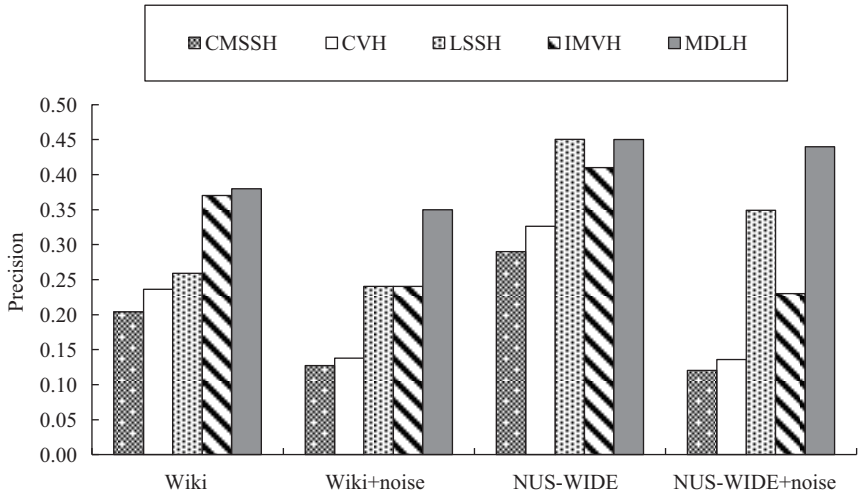

Figure 10 Mean average precisions of different methods with and without noise.

include many words irrelevant to the topic, and second, tags provide more semantic information than the words in the paragraphs.

4.4.3 *Results on noisy datasets*

To evaluate the robustness to noise of each method, we add noises to the Wikipedia–Picture of the Day and NUS-WIDE datasets separately, and compare the performances on the noisy datasets. For both datasets, we select a category randomly as the source of the respective noise. Some pictures and words from them are selected randomly as noise to be added to the rest of the data. In the Wikipedia–Picture of the Day dataset, we select 2% of the text and one picture as noise each time. In the NUS-WIDE dataset, we select one tag as the noise. Figure 10 shows the performance before and after the addition of noise. Compared with the results on the original datasets, the performance on the noisy datasets of all the methods decreases to some degree, with the mAP of MDLH decreasing less than that of the other methods. This shows that our method is more robust to noise than the others.

4.4.4 *Influence of unlabeled dataset*

In real-world applications, new data keep appearing in the database as time goes by. The amount of such data can be so great that it is impossible to annotate them owing to the cost in time. The ability to deal with such unlabeled data can enhance the applicability of hashing methods to practical problems. As mentioned in Subsection 3.2, MDLH use unlabeled data to pre-train the stacked autoencoder for each modality and learn the common representation. To evaluate the influence of unlabeled data, we apply the algorithm in two settings: using unlabeled data (MDLH) and using only labeled training data

Table 5 MAP of two training settings on the Wikipedia–Picture of the Day and NUS-WIDE datasets

Dataset	Task	Method	HCL = 16	HCL = 32	HCL = 64
Wikipedia–Picture of the Day	Image query text	MDLH-TD	0.4501	0.4403	0.4546
		MDLH	**0.4526**	**0.4537**	**0.4555**
	Text query image	MDLH-TD	0.4370	0.4290	0.4301
		MDLH	**0.4496**	**0.4478**	**0.4485**
NUS-WIDE	Image query text	MLDH-TD	0.3827	0.3912	0.3930
		MDLH	**0.3919**	**0.3940**	**0.4030**
	Text query image	MDLH-TD	0.3785	0.3653	0.3529
		MDLH	**0.3840**	**0.3729**	**0.3604**

(MDLH-TD). Table 5 shows the mAP of these two settings. From the table, we can see that unlabeled data improve the performance. Because the number of unlabeled data is small compared with training data, the improvement is not very significant. In the future, we shall consider using large numbers of unlabeled data from different datasets.

5 Conclusion

In this paper, we have proposed a multimodal deep-learning-based cross-modal hash learning method. The multimodal deep learning is used to model the relationship between multiple heterogeneous data and learn a shared representation of the multimodal data, which is robust to noise and easy to extend to multiple modalities. Experiments on two realistic datasets show that our method represents the multimodal features effectively.

However, the proposed multimodal deep learning structure models the cross-modal relation entirely on the joint layer, which ignores the relationship between each pair of modalities. Moreover, the feature learning process for individual modalities could be improved by using multiple features for each modality.

In the future, we will focus on improving the algorithm in two respects: (1) We shall extend the model to media types such as audio and video. According to their individual characteristics, we shall consider more features for each modality; (2) We shall consider multiple relationships between modalities by adding hidden layers to model pairwise relationships.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61402091, 61370074), and Fundamental Research Funds for the Central Universities of China (Grant No. N140404012).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Chen C, Zhu Q S, Lin L, et al. Web media semantic concept retrieval via tag removal and model fusion. ACM Trans Intel Syst Technol, 2013, 4: 478–488
2 Leung C H C, Chan A W S, Milani A, et al. Intelligent social media indexing and sharing using an adaptive indexing search engine. ACM Trans Intel Syst Technol, 2012, 3: 338–343
3 Zhang R M, Lin L, Zhang R, et al. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Trans Imag Process, 2015, 24: 4766–4779
4 Nie X S, Liu J, Sun J D, et al. Robust video hashing based on representative-dispersive frames. Sci China Inf Sci, 2013, 56: 068104
5 Xiang S J, Yang J Q, Huang J W. Perceptual video hashing robust against geometric distortions. Sci China Inf Sci, 2012, 55: 1520–1527
6 Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of ACM Symposium on Computational Geometry, New York, 2004. 253–262
7 Weiss Y, Torralba A, Fergus R. Spectral hashing. In: Proceedings of 22nd Annual Conference on Neural Information Processing Systems, Vancouver, 2008. 1753–1760

8 Zhen Y, Yang D. A probabilistic model for multimodal hash function learning. In: Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, 2012. 940–948
9 Zhu X F, Huang Z, Shen H T, et al. Linear cross-modal hashing for efficient multimedia search. In: Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, 2013. 143–152
10 Yu Z, Wu F, Yang Y, et al. Discriminative coupled dictionary hashing for fast cross-media retrieval. In: Proceedings of the 37th Annual ACM SIGIR Conference, Gold Coast, 2014. 395–404
11 Bronstein M, Bronstein A, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 3594–3601
12 Kumar S, Udupa R. Learning hash functions for cross-view similarity search. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, 2011. 1360–1365
13 Hu Y, Jin Z M, Ren H Y, et al. Iterative multi-view hashing for cross media indexing. In: Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, 2014. 527–536
14 Song J K, Yang Y, Yang Y, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, New York, 2013. 785–796
15 Wu B T, Yang Q, Zheng W S, et al. Quantized correlation hashing for fast cross-modal search. In: Proceedings of International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 3946–3952
16 Kang Y, Kim S, Choi S. Deep learning to hash with multiple representations. In: Proceedings of IEEE International Conference on Data Mining, Brusselsm, 2012. 930–935
17 Wang D X, Cui P, Ou M D, et al. Deep multimodal hashing with orthogonal regularization. In: Proceedings of International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 2291–2297
18 Wang Q F, Si L, Shen B. Learning to hash on partial multimodal data. In: Proceedings of International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 3904–3910
19 Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech, 2012, 20: 30–42
20 Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: Proceedings of Annual Conference on Neural Information Processing Systems, Lake Tahoe, 2012. 1106–1114
21 Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: Proceedings of International Conference on Machine Learning, Washington, 2011. 689–696
22 Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, 2012. 2231–2239
23 Sohn K, Shang W, Lee H. Improved multimodal deep learning with variation of information. In: Proceedings of the 28th Annual Conference on Neural Information Processing Systems, Montreal, 2014. 2141–2149
24 Wu P C, Hoi S C, Xia H, et al. Online multimodal deep similarity learning with application to image retrieval. In: Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, 2013. 153–162
25 Wang W, Ooi B C, Yang X Y, et al. Effective multi-modal retrieval based on stacked autoencoders. In: Proceedings of 40th International Conference on Very Large Data Bases, Hangzhou, 2014. 649–660
26 Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 21st ACM International Conference on Multimedia, Orlando, 2014. 7–16
27 Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res, 2010, 11: 3371–3408
28 Salakhutdinov R, Hinton G. Deep Boltzmann machines. In: Proceedings of 12th International Conference on Artificial Intelligence and Statistics, Florida, 2009. 448–455
29 Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
30 Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks. In: Proceedings of Annual Conference on Neural Information Processing Systems, Vancouver, 2006. 153–160
31 Rumelhart D, Hinton G, Williams R. Neurocomputing: Foundations of Research. Cambridge: MIT Press, 1988
32 Rasiwasia N, Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, New York, 2010. 251–260
33 Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
34 Lowe D. Distinctive image features from scale-invariant key points. Int J Comput Vision, 2004, 60: 91–110
35 Chua T S, Tang J H, Hong R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of ACM International Conference on Image and Video Retrieval, Santorini, 2009. 1–9
36 Zhou J, Ding G G, Guo Y C. Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th Annual International ACMSIGIR Conference, Gold Coast, 2014. 415–424