



# Hybrid representation learning for cross-modal retrieval

Wenming Cao<sup>a,b</sup>, Qiubin Lin<sup>a,\*</sup>, Zhihai He<sup>b</sup>, Zhiquan He<sup>a</sup>

<sup>a</sup>Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China

<sup>b</sup>Video Processing and Communication Laboratory, Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211, USA

## ARTICLE INFO

### Article history:

Received 7 June 2018

Revised 30 September 2018

Accepted 8 October 2018

Available online 4 February 2019

### Keywords:

Cross-modal retrieval

Hybrid representation

DNNs

## ABSTRACT

The rapid development of Deep Neural Networks (DNNs) in single-modal retrieval has promoted the wide application of DNNs in cross-modal retrieval tasks. Therefore, we propose a DNN-based method to learn the shared representation for each modality. Our method, hybrid representation learning (HRL), consists of three steps. In the first learning step, stacked restricted Boltzmann machines (SRBM) are utilized to extract the modality-friendly representation for each modality, with statistical properties that are more similar than those of the original input instances of both modalities, and a multimodal deep belief net (multimodal DBN) is utilized to extract the modality-mutual representation, which contains some missing information in the original input instances. In the second learning step, a two-level network containing a joint autoencoder and a three-layer feedforward neural net are used. From these steps, the hybrid representation is obtained, which combines the image representation constructed by the image-pathway SRBM and the modality-mutual representation, which involves the latent image representation and can be used to infer the missing values of the image via the multimodal DBN or vice-versa. In the third learning step, stacked bimodal autoencoders are used to obtain the final shared representation for each modality. The experimental results show that our proposed HRL method is superior to several advanced approaches according to three widely used cross-modal datasets.

© 2019 Elsevier B.V. All rights reserved.

## 1. Introduction

With the rapid development of technology, the world around us consists of increasingly more multimodal data. Images depicted by text and videos containing visual and audio signals are examples of multimodal data. The presence of large amounts of multimodal data places high demands on cross-modal retrieval [1,2], such as using text to retrieve images and using an image to retrieve text. Unlike traditional retrieval methods in a single modality (i.e., image retrieval [3] and text retrieval), the task of cross-modal retrieval focuses on how to use correlations between various modalities.

Cross-modal retrieval is a fairly new retrieval method, which can be utilized to retrieve multimodal data. For example, if someone is interested in a singer, he can use an image query and then obtain relevant multimodal data including images, text descriptions, videos, audio clips and so on. An example of cross-modal retrieval using an image and text is illustrated in Fig. 1. Compared with single-modal retrieval, cross-modal retrieval can

provide users with a more flexible and useful search experience because it can display rich multimodal retrieval results. However, the challenge of cross-modal retrieval is that the representations of text and images fail to directly match with each other because the distribution and representation of different modalities are inconsistent.

### 1.1. Previous work

To solve the above-mentioned challenge, many methods have been proposed. The traditional cross-modal retrieval methods rely mainly on learning the common representation space of various modalities. A typical method is canonical correlation analysis (CCA) [4], which builds the representation subspace to mine the correlation between various modalities and is extensively utilized by [5–7] to build multimodal data. For example, Rasiwaisa et al. [8] managed to associate CCA with semantic category information and multiview CCA [9] proposed by Gong et al., incorporates the third view of high-level semantic information with CCA. Similar to CCA, another method is cross-modal factor analysis (CFA) [10], which minimizes the Frobenius norm in the common representation subspace to find projection functions for various modalities. The multi-label CCA method proposed by Ranjan et al. [11] is an extension of CCA, which takes into account high-level semantic

\* Corresponding author.

E-mail addresses: [wmcao@szu.edu.cn](mailto:wmcao@szu.edu.cn) (W. Cao), [linqiubin2017@email.szu.edu.cn](mailto:linqiubin2017@email.szu.edu.cn), [2170269126@email.szu.edu.cn](mailto:2170269126@email.szu.edu.cn) (Q. Lin), [heZhi@missouri.edu](mailto:heZhi@missouri.edu) (Z. He), [zhiquan@szu.edu.cn](mailto:zhiquan@szu.edu.cn) (Z. He).

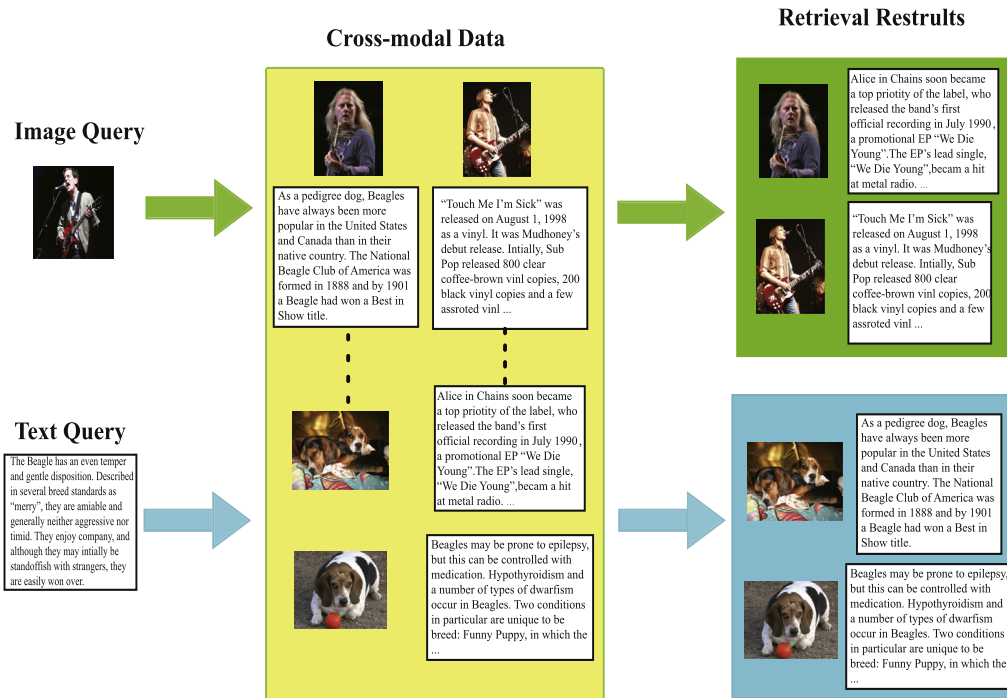


Fig. 1. An example of cross-modal retrieval using an image and text that presents retrieval results with various modalities.

information in the form of multi-label annotations. Tran et al. [12] proposed a new representation method, namely, Multimedia Aggregated Correlated Components (MACC), which aggregates information provided by the projections of the image and text representation on their aligned subspaces. Moreover, TINA [13], a cross-modal correlation method that employs adaptive hierarchical semantic aggregation, was put forward to construct a set of local projections and gating functions for image and text.

Recently, semi-supervised learning and graph regularization have been utilized in some methods to learn cross-modal representation. For example, Zhai et al. proposed joint graph regularized heterogeneous metric learning (JGRHML) [14] to integrate the structure of various modalities into a joint graph regularization. Joint representation learning (JRL) [15] utilizes semantic information with semi-supervised and sparse regularization for various modalities to construct the correlation in a unified framework. Wang et al. [16] learned projection matrices to map various modal data into a common subspace with an iterative algorithm, which intends to retain intermodality and intramodality similar relations.

Although the above traditional cross-modal retrieval methods have made great progress, they still cannot fully exploit the inherent correlation of cross-modal data because they rely mainly on a linear projection.

In recent years, the rapid development of Deep Neural Networks (DNNs) [17] in single-modal retrieval and classification (e.g. text/image/video classification [18–20]) has promoted wide application of DNNs in cross-modal retrieval tasks. DNNs have been used to mine the correlation of cross-modal data by converting cross-modal data to shared representations [21–23]. The bimodal autoencoder (bimodal AE) [24] is used for a cross-modal retrieval task, and various modal inputs are passed through the network to obtain the final shared representation. Based on this idea, several similar architectures were later proposed, and progress has been made in building the cross-modal representation [25–27].

Most of these DNNs-based methods consist of two steps. The first step is to extract the respective representation of each modality. Then, the second step builds the shared representation by

mining the correlation between two modalities. In the first step, some methods only extract intramodality representations, while others simultaneously extract intramodality and intermodality representations. Although intramodality representation reveals intrinsic properties of each modality, this leads to a further distance in the common representation space. In addition, although intermodality representation can be used to exploit the inherent correlation between two modalities, it ignores the missing mutual information in the original input representation of each modality. In the second step, existing methods [28,29] use a shallow network architecture to learn the shared representation, which cannot fully exploit the complex cross-modality correlation.

## 1.2. Motivation and contribution

As illustrated in Fig. 2, some given tags cannot fully describe all details about an image; however, some generated tags can offer some complementary details about the image that are missing in the original input representation of the text. Motivated by this phenomenon, we adopt a multimodal deep belief network (multimodal DBN) [28], which can generate text and image representations, to acquire more useful information about input instances.

In this paper, hybrid representation learning (HRL) is proposed to mine the rich and complex cross-modality correlation. The main contributions of our work can be summarized as follows.

- We propose a novel framework which can fully consider and utilize missing information in original input instances for each modality.
- Hybrid representation combines image representation, which is constructed by the image pathway stacked restricted Boltzmann machines (SRBM), and text representation, which involves the latent image representation and can be used to infer missing values of the image through a multimodal deep belief network (multimodal DBN), or vice-versa.
- Experimental results on three widely used datasets show that our proposed HRL method outperforms the state-of-the-art methods.


| Image   | Given Tags   | Generated Tags   |
|---|--|--|
|  | pentax, k10d,<br>kangarooisland,<br>southaustralia,<br>sa, australia,<br>australiansealion,<br>300mm | beach, sea,<br>surf, strand,<br>shore, wave,<br>seaspace,<br>sand, ocean,<br>waves                       |
|  | mickikimmel,<br>mickipeida,<br>headshot  | portait, girl,<br>woman, lady,<br>blone, pretty,<br>gorgeous,<br>expression,<br>model                    |
|  | camera,<br>lightpainting,<br>reflection,<br>doublepaneglass,<br>wowiekazowie,<br>jahdakine           | blue, art,<br>artwork,<br>artistic, surreal,<br>expression,<br>original, artist,<br>gallery,<br>patterns |

Fig. 2. Examples of data from the MIR Flickr Dataset [30], along with text generated from the Deep Belief Net [28] by sampling from  $P(v_{\text{txt}}|v_{\text{img}}, \theta)$ .

The rest of the paper is organized as follows. In Section 2, some related existing work will be introduced. Then, in Section 3, our proposed HRL method is described in detail and evaluated experimentally in Section 4. Finally, in Section 5, we draw our conclusions for the paper.

## 2. Related work

Deep neural networks have presented powerful capabilities in mining nonlinear correlation and have performed excellently in some single-modal tasks, such as text/image/video classification [18–20,31–33] and object detections [34–37]. Motivated by this, some general models are applied to represent various modalities in our HRL method. Considering our innovative hybrid representation learning, the restricted Boltzmann machine (RBM) and the multimodal DBN are utilized as the building blocks in our HRL method. In particular, the modality-friendly and modality-mutual representations are constructed by SRBM and a multimodal DBN, respectively, in our HRL method. Then, modality-friendly and modality-mutual representations are combined by joint autoencoders to generate a hybrid representation, and a bimodal AE in a stacked style is applied to mine the correlation between hybrid representations of two modalities. These basic models will be briefly introduced as follows.

### 2.1. Restricted Boltzmann machine

The RBM [38] is an undirected graphical model, typically with stochastic binary units in the visible layer and the hidden layer; however, any two units in the same layer are not connected. Its probabilistic distribution is defined as follows:

$$P(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta)), \quad (1)$$

where  $Z(\theta)$  is the normalizing constant.  $E(\mathbf{v}, \mathbf{h}; \theta)$  is the energy function defined as follows:

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_{i=1}^n \sum_{j=1}^m w_{ij} v_i h_j - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j, \quad (2)$$

where  $v \in \{0, 1\}^D$  are visible units,  $h \in \{0, 1\}^F$  are hidden units, and  $\theta = \{a, b, w\}$  are the model parameters. It can be easily derived from Eqs. (1) and (2) that the conditional distributions over the visible units  $\mathbf{v}$  and hidden units  $\mathbf{h}$  are given as:

$$p(h_j = 1 | \mathbf{v}) = \sigma \left( b_j + \sum_{i=1}^n w_{ij} v_i \right) \quad (3)$$

$$p(v_i = 1 | \mathbf{h}) = \sigma \left( a_i + \sum_{j=1}^m w_{ij} h_j \right) \quad (4)$$

RBM can achieve the modality-friendly representation whose statistical properties are more similar than those of the original input instances of both modalities.

### 2.2. Modeling real-valued data: Gaussian RBM

The Gaussian RBM [39] is a variant of RBM, which can be utilized to model real-valued data by using a Gaussian distribution instead of a Bernoulli distribution. The energy of the Gaussian RBM is defined as:

$$E(\mathbf{v}, \mathbf{h}; \theta) = \sum_{i=1}^n \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^n \sum_{j=1}^m w_{ij} \frac{v_i}{\sigma_i} h_j - \sum_{j=1}^m b_j h_j, \quad (5)$$

where  $\theta = \{a, b, w\}$  are the model parameters.

### 2.3. Modeling count data: replicated softmax RBM

The replicated softmax RBM [40] is utilized to model the sparse bag of words vectors in a document. Given a document which contains  $D$  words, the energy function of the replicated softmax RBM is defined as:

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_{i=1}^n \sum_{j=1}^m w_{ij} v_i h_j - \sum_{i=1}^n a_i v_i - D \sum_{j=1}^m b_j h_j, \quad (6)$$

where  $\theta = \{a, b, w\}$  are the model parameters.

### 2.4. Multimodal deep belief network

As illustrated in Fig. 3(a), a multimodal DBN [28] is proposed to learn the common representation of various modalities. It uses a separate two-layer DBN to construct the distribution of original representation, where the image representation is learned through the Gaussian RBM, and the text representation is learned through the replicated softmax RBM. The probability which each specific DBN assigns to a visible vector is:

$$P(v_i) = \sum_{h_i^{(1)}, h_i^{(2)}} P(h_i^{(2)}, h_i^{(1)}) P(v_i | h_i^{(1)}) \quad (7)$$

$$P(v_t) = \sum_{h_t^{(1)}, h_t^{(2)}} P(h_t^{(2)}, h_t^{(1)}) P(v_t | h_t^{(1)}) \quad (8)$$

To form a multimodal DBN, these two specific DBN are combined by a joint RBN on the top of them. The joint distribution can be written as follows:

$$\begin{aligned} P(v_i, v_t) = & \sum_{h_i^{(2)}, h_t^{(2)}, h^{(3)}} P(h_i^{(2)}, h_t^{(2)}, h^{(3)}) \\ & \times \sum_{h_i^{(1)}} P(v_i | h_i^{(1)}) P(h_i^{(1)} | h_i^{(2)}) \\ & \times \sum_{h_t^{(1)}} P(v_t | h_t^{(1)}) P(h_t^{(1)} | h_t^{(2)}) \end{aligned} \quad (9)$$

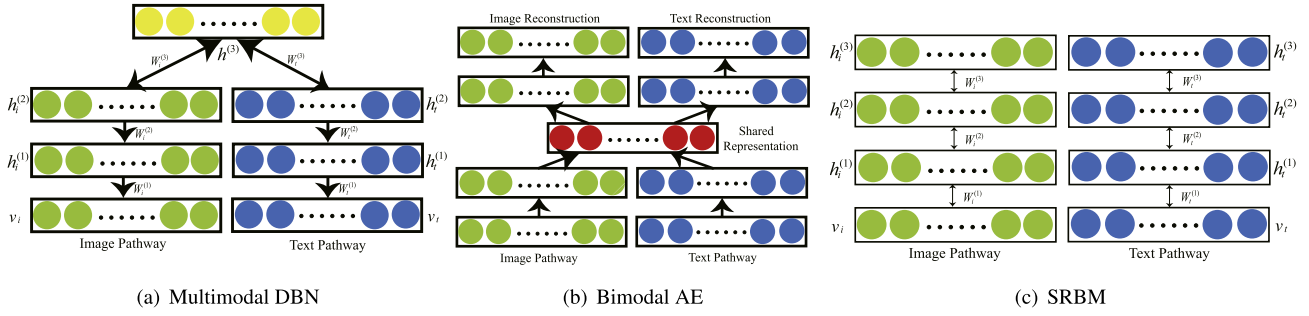


Fig. 3. Basic models used in our HRL method.

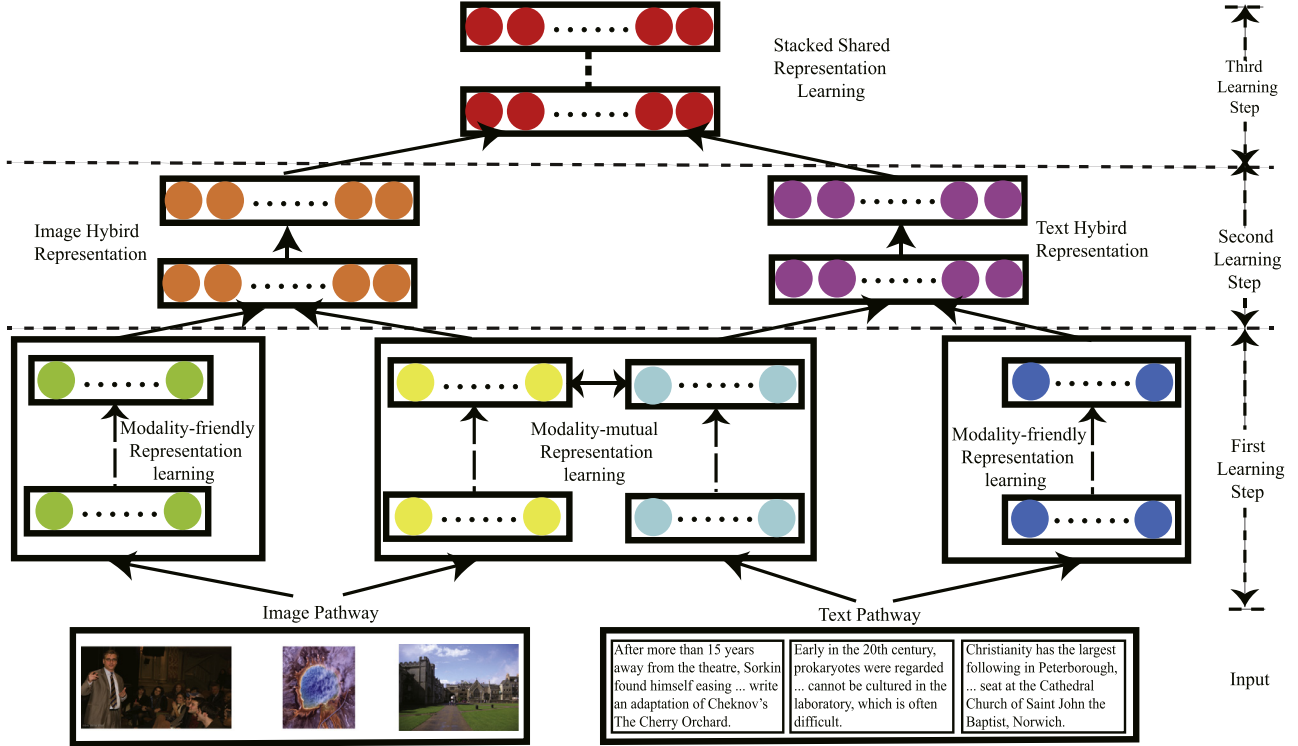


Fig. 4. An overview of our HRL method.

### 2.5. Bimodal autoencoder

As illustrated in Fig. 3(b), the bimodal AE [24] is an extension of RBM for constructing various modalities. It utilizes different modalities as inputs and generates a common representation in the shared joint layer. It can reconstruct various modalities such as text and images by minimizing the error between the original input representation and the final reconstructed representation. It can also retain the reconstruction information in each reconstructed modality representation by exploiting the high-order correlation between various modalities.

## 3. Our proposed method

In this section, we will introduce our proposed method HRL in detail. As shown in Fig. 4, our HRL method includes three steps. In the first learning step, HRL utilizes SRBM to learn the modality-friendly representation for each modality and applies the multimodal DBN to learn the modality-mutual representation. In the second learning step, HRL uses a two-level network containing a joint auto-encoder and a three-layer feedforward neural net to obtain the hybrid representation. In the third learning

step, HRL applies stacked bimodal AEs to obtain the final shared representation for each modality.

First, the formal definition will be given as follows. The multimodal dataset includes two modalities with  $m$  image cases and  $n$  text cases, which is denoted as  $D = \{D^{(i)}, D^{(t)}\}$ . Here,  $D^{(i)} = \{x_p^{(i)}, y_p^{(i)}\}_{p=1}^m$  denotes the image item, and  $D^{(t)} = \{x_q^{(t)}, y_q^{(t)}\}_{q=1}^n$  denotes the text case, where  $x_p^{(i)} \in \mathbb{R}^{d^{(i)}}$  denotes the  $p$ th image case and  $x_q^{(t)} \in \mathbb{R}^{d^{(t)}}$  denotes the  $q$ th text case, which are labeled as  $y_p^{(i)}$  and  $y_q^{(t)}$ , respectively. In addition,  $d^{(i)}$  and  $d^{(t)}$  are the dimension of the image and text representation, respectively.

### 3.1. Modality-friendly and modality-mutual representation learning

In our HRL method, the first step contains modality-friendly and modality-mutual representation learning.

**Modality-friendly representation learning.** As shown in Fig. 3(c), SRBM are utilized to learn the modality-friendly representation. In the first two layers of the modality-friendly representation learning step, SRBM applies the Gaussian RBM to extract the real-valued representation for an image and uses the replicated softmax RBM to extract the sparse word count vectors for text. After learning



modality-specific RBMs, two basic RBMs are applied to remove modality-specific properties to obtain the high-level semantic representation. In this step, the probability which each modality-friendly SRBM assigns to an output vector is:

$$P(h_i^{(3)}) = \sum_{v_i, h_i^{(1)}, h_i^{(2)}} P(h_i^{(3)} | h_i^{(2)}) P(h_i^{(2)} | h_i^{(1)}) P(h_i^{(1)} | v_i) \quad (10)$$

$$P(h_t^{(3)}) = \sum_{v_t, h_t^{(1)}, h_t^{(2)}} P(h_t^{(3)} | h_t^{(2)}) P(h_t^{(2)} | h_t^{(1)}) P(h_t^{(1)} | v_t) \quad (11)$$

After training the SRBM, we can obtain the modality-friendly representation  $X_{friendly}^i$  and  $X_{friendly}^t$ , whose statistical properties are more similar than those of the original input instances of both modalities.

**Modality-mutual representation learning.** In this step, a multimodal DBN is applied to learn the modality-mutual representation. First, the image-specific DBN makes use of the Gaussian RBM to construct the distribution over real-valued image representations, while the text-specific DBN takes advantage of the replicated softmax RBM to model the distribution over sparse word count vectors. Then, these two separate modality-specific DBNs are combined by a joint RBM on the top of them to learn the modality-mutual representation. As shown in Fig. 3(a), for example, we consider generating an image conditioned on given text  $v_t$ .<sup>1</sup> According to [28], alternating Gibbs sampling is performed through the joint layer using the following conditional distributions:

$$P(h_t^{(3)} | h_t^{(2)}, h_i^{(2)}) = \sigma(W_t^{(3)} h_t^{(2)} + W_i^{(3)} h_i^{(2)} + a_t) \quad (12)$$

$$P(h_i^{(2)} | h_t^{(3)}) = \sigma(W_i^{(3)} h_t^{(3)} + a_i) \quad (13)$$

where  $\sigma(x) = 1/(1 + e^{-x})$ . We denotes  $h_t^{(3)}$  as  $X_{mutual}^i$ , which apparently contains the text representation, but it actually involves the latent image representation which can generate the image case  $h_i^{(2)}$ . Therefore, we can use  $X_{mutual}^i$  as the image representation, which can be utilized to infer missing values of the image through the multimodal DBN. Similarly,  $X_{mutual}^t$  is used to denote the text representation.

### 3.2. Hybrid representation learning with a two-level network

In the second learning step, the diverse, complementary and separate representations  $X_{friendly}^i$ ,  $X_{friendly}^t$ ,  $X_{mutual}^i$  and  $X_{mutual}^t$  for each modality have been acquired, which can stand for both the modality-friendly and modality-mutual representations in the first learning step. To generate the hybrid representation,  $X_{friendly}^i$ , which stands for image representation in the image-pathway SRBM, and  $X_{mutual}^i$ , which originally embodied text representation and can infer missing image values via the multimodal DBN, are combined by a learning method containing a deeper two-level networks.

On the first level of the network, a joint autoencoder is used to combine modality-friendly and modality-mutual representations of each modality. It jointly models the distribution over the representations obtained from the SRBM and multimodal DBN. The joint distribution can be defined as follows:

$$P(v_1, v_2) = \sum_{h_1^{(1)}, h_2^{(1)}, h^{(2)}} P(h_1^{(1)}, h_2^{(1)}, h^{(2)}) \sum_{h_1^{(1)}} P(v_1 | h_1^{(1)}) \times \sum_{h_2^{(1)}} P(v_2 | h_2^{(1)}), \quad (14)$$

where  $v_1$  symbolizes the modality-friendly representation  $X_{friendly}^t$  for text, and  $v_2$  denotes the modality-mutual representation  $X_{mutual}^t$  for text. In addition, for the image modality, this joint distribution is applied on the modality-friendly representation  $X_{friendly}^i$  and the modality-mutual representation  $X_{mutual}^i$  for the image. These joint distributions are collected as the intermediate hybrid representations for each kind of modality, which are symbolized as  $Y^{(t)}$  for text and  $Y^{(i)}$  for image. In addition, they will be utilized as the input of the next step in the network.

On the second level of the network, a three-layer feedforward neural net with a softmax layer is used to mine useful category information of these hybrid representations.

### 3.3. Hybrid representation correlation learning

In the third learning step, several bimodal AEs are utilized to mine the correlation of the hybrid representations generated by the last learning step. Several bimodal AEs can reconstruct various modalities such as text and images by minimizing the reconstruction error between the original input features and the final reconstructed representation. For the purpose of training the network, a stacked learning method has  $n$  bimodal AE trained in a bottom-up method following [24]. The intermediate representation  $Y^{(i)}$  and  $Y^{(t)}$  are utilized as input for the bottom bimodal AE, and its output  $Z_1^{(i)}$  and  $Z_1^{(t)}$  will be applied as the input to propagate to the higher network in order to obtain  $Z_2^{(i)}$  and  $Z_2^{(t)}$  as output which reduces the dimension to half of the input at the same time, until the final shared representation  $Z_n^{(i)}$  and  $Z_n^{(t)}$  are acquired. The parameter  $n$ , which symbolizes the number of the bimodal AE to be stacked in the training process, can be adjusted dynamically according to the validation set. Through  $n$  stacked bimodal AEs which have performed better than only one bimodal AE, the final shared representation can be obtained. Therefore, the modality-friendly and modality-mutual information can be jointly constructed to mine the complex and rich cross-modal correlation.

## 4. Experiment

This section will introduce our experiments on three widely used cross-modal datasets (Wikipedia, NUS-WIDE-10k and Pascal Sentences dataset) with eleven state-of-the-art methods. To fully and objectively evaluate the results, two kinds of retrieval tasks are conducted, which are bi-modal retrieval and all-modal retrieval. To further verify the validity of our HRL method, we also conduct experiments with joint representation instead of hybrid representation and other baseline experiments.

### 4.1. Experimental setup

The three above-mentioned datasets are briefly introduced here. For a fair comparison, in our experiments, feature extraction strategies and dataset partition are strictly conducted according to [29,41], and are also exactly the same in all the comparison methods.

**Wikipedia dataset** [8]. The dataset contains 2866 image/text pairs and is randomly separated into 2173 cases for training, 231 cases for validation and 462 cases for testing. The dimension of the text feature is 3000, which is extracted by a bag of words model. For image representation, it is concatenated by three components: 1000-dimensional pyramid histogram of words, 512-dimensional Gist descriptors and a 784-dimensional MPGE-7 feature vector.

**NUS-WIDE-10k dataset** [42]. The dataset is a subset of the NUS-WIDE dataset which consists of approximately 270,000 images and is divided into 81 categories. NUS-WIDE-10k is built by selecting 10,000 image/text pairs evenly from the 10 largest categories.

<sup>1</sup> Generating text representation conditioned on an image can be done in similar way.

**Table 1**  
Statistics of the datasets used in this paper.

| Statistics                 | Wikipedia | NUS-WIDE-10k | Pascal Sentences |
|----------------------------|-----------|--------------|------------------|
| Number of training         | 2173      | 8000         | 800              |
| Number of validation       | 231       | 1000         | 100              |
| Number of test data        | 462       | 1000         | 100              |
| Number of categories       | 10        | 10           | 20               |
| Image hand-crafted feature | 2296      | 1134         | 2296             |
| Image CNN feature          | 4096      | 4096         | 4096             |
| Text modality              | article   | tags         | sentences        |
| Text feature               | 3000      | 1000         | 1000             |

According to [29], the dataset is split into three subsets: 8,000 instances for training set (800 instances per category), 1,000 instances for the validation set (100 instances per category) and 1000 instances for the test set (100 instances per category). The text representation is embodied by a 1000-D bag of words vector. The image representation is concatenated by a 64-D color histogram, a 144-D color correlogram, a 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D SIFT-based bag-of-words features.

*Pascal Sentences dataset* [43]. The dataset contains 1000 image/text pairs of 20 categories (50 cases per class), which is selected from the 2008 PASCAL development kit. Every image is tagged with five sentences. By random selection, 800 are used for training set (40 cases per category), 100 for validation set (5 cases per category) and 100 for testing set (5 cases per category). The image representation extraction method is the identical to that used with the Wikipedia dataset. In addition, the text feature is a 1000-dimensional bag of words vector.

Moreover, since convolution neural network (CNN) [44] features have been proposed for their effectiveness on image features, CNN image representations are used in our experiments. Specifically, for all comparison methods on three datasets, the CNN features used are represented with 4,096 dimensions extracted from the fc7 layer of VGGNet [45].

As shown in Table 1, these datasets have very different properties. For example, the text modality of the three datasets Wikipedia, NUS-WIDE-10k and Pascal Sentences is completely different, namely, an article, tags and sentences, respectively. In addition, these datasets range in size from 1k to 10k, with categories ranging from 10 to 20.

#### 4.2. Details of the HRL method

Here, we describe our method in detail. The HRL method is implemented based on deepnet.<sup>2</sup> As illustrated in Fig. 4, the HRL method has three major components, none of which is dedicated to a single modality. After extracting the representation, any modality can be embodied as a representation vector and can be used as the HRL's input for cross-modal retrieval.

First, HRL utilizes the SRBM with three layers to learn the separate modality-friendly representation. For the first two layers of each modality, the Gaussian RBM and replicated softmax RBM are used to learn modality-specific representations, both of which contain 1024 hidden units on the first layer and 512 on the second layer. After learning the modality-specific RBMs, the hidden layer can be utilized as inputs of the basic RBMs, which are the same for both special modalities with hidden units of 512 dimensions.

Second, HRL uses the multimodal DBN to learn the modality-mutual representation. For image modality, the multimodal DBN uses a two-layer DBN that contains 2048-dimensional hidden units on the first layer and 1024-dimensional hidden units on the second

layer. For text modality, a two-layer DBN is utilized, which includes 1024-dimensional hidden units on the first two layers. The joint layer uses the outputs of two separate modality DBNs as inputs and utilizes a joint RBM with 2048-dimensional hidden units on the top of inputs.

Third, HRL uses a two-level network to learn the hybrid representation. At the first level, a joint autoencoder containing 1024-dimensional hidden units is utilized to combine the modality-friendly and modal-mutual representations. At the second level, a three-layer feedforward neural network is used on the top of each joint AE, which has 1024-dimensional units at each layer and a softmax layer at the last layer.

Finally, HRL uses the bimodal AE to obtain the final shared representation. The reconstruction layers have the identical dimensions as the input. The dimension of the joint layer is half of the input, from which the final shared representation can be obtained. Additionally, a softmax layer is also used to connect the joint layer for further optimization. As described in Section 3, the amount of bimodal AEs is adjusted based on the validation set.

In the training step, image and text input cases should be combined into pairs, but in the test step, they are actually independent of each other. Furthermore, the dimension of the above-mentioned input is exemplified by the NUS-WIDE-10k, which has a 1000-dimensional text representation and a 1134-dimensional image representation, and they require adjustments based on input dimensions of other datasets. In addition to the above details, other parameters are consistent with implementations of the existing network deepnet.<sup>2</sup>

#### 4.3. Evaluation metrics

Two kinds of cross-modal retrieval tasks are conducted on the Wikipedia, NUS-WIDE-10k and Pascal Sentences datasets following [29,47]:

- *Bi-modal retrieval.* Utilizing another modality to retrieve one modality in the dataset, that is, searching for text by an image (image  $\rightarrow$  text) and retrieving images by text (text  $\rightarrow$  image).
- *All-modal retrieval.* All modalities in the dataset are retrieved using any modality, that is, retrieving both images and text by an image (image  $\rightarrow$  all) and text(text  $\rightarrow$  all).

In this paper, the mean average precision (MAP) score and precision-recall (PR) curve are utilized to evaluate the retrieval performance on the Wikipedia, NUS-WIDE-10k, Pascal Sentences datasets following [41]. Given a set of queries, the MAP score refers to the mean of average precision (AP) for each query. AP is defined as follows:

$$AP = \frac{1}{R} \sum_{k=1}^n \frac{R_k}{k} \times rel_k \quad (15)$$

where  $R$  means the number of relevant items,  $n$  is the number of instances in retrieval set and  $R_k$  counts the number of relevant items in the top  $k$  results. If the  $k$ -th result is relevant,  $rel_k$  is set to be 1; otherwise, 0. It should be noted that our experimental results of the MAP score are on all returned results, which is widely utilized in cross-modal retrieval tasks such as [8,41,47], and does not adopt only the top 50 as [29].

#### 4.4. Comparison with state-of-the-art methods

We compare our method with eleven state-of-the-art cross-modal retrieval methods, namely, CCA [4], CFA [10], KCCA (poly) [5], KCCA (Gaussian) [5], multimodal DBN [28], bimodal AE [24], corr-AE [29], JRL [15], CMDN [47], ACMR [46], and CCL [41]. CCA, KCCA (poly), KCCA (Gaussian) and CFA are classical baselines. Multimodal DBN, bimodal AE and corr-AE are cross-modal retrieval

<sup>2</sup> <https://github.com/nitishsrivastava/deepnet>.

methods based on DNNs. JRL is the most advanced method based on a linear projection. Both CMDN and CCL mine intramodal and intermodal correlation to improve their performance. ACMR utilizes adversarial learning to find a common subspace. Additionally, comprehensive evaluations of MAP use all retrieval results instead of the top 50 results as [29]. The eleven comparison methods are introduced concisely as follows.

- *CCA* is a method of learning linear relationship between two modalities. It learns mapping matrices to project the representations of various modalities into a lower-dimensional common space, and then the shared representation is obtained.
- *CFA* projects the cross-modal data into the common representation space by learning a linear projection, which minimizes the Frobenius norm between both modalities.
- *KCCA* first projects the original features into a higher-dimensional representation space before performing CCA in the new representation space. The kernel functions used in our experiments are the polynomial (poly) kernel and Gaussian kernel.
- *Multimodal DBN* models the distribution of the original features for each modality by training the two specific two-layer DBNs in the first step. Then, in the second step, it utilizes a joint RBM on the top of the two specific DBNs and combines them by modeling the joint distribution of the data with various modalities to obtain the final shared representation.
- *Bimodal AE* is a kind of deep autoencoder network that can reconstruct various modalities such as image and text jointly by minimizing the reconstruction error between the original input features and the final reconstructed representations. It can learn a high-order correlation between various modalities and preserve the reconstruction information at the same time.
- *Corr-AE* combines representation learning and correlation learning into a single process while constructing the correlation loss and the reconstruction error through two subnetworks connected to the code layers. Although there are two extension of corr-AE: corr-cross-AE and corr-full-AE, we use the model with the best MAP score among the three models in our experiment.
- *JRL* aims to learn a sparse projection matrix, so that the original heterogeneous features can be directly projected into a joint space. It integrates the sparse and semi-supervised regularization of all modalities into a unified optimization framework to jointly mine the correlation and semantic information.
- *CMDN* jointly mines the intramodal and intermodal information to obtain the complementary separate representation of each modality in the first step. In the second step, it hierarchically combines the intermodal and intramodal representations to further exploit the rich cross-modal correlation.
- *ACMR* is based on an adversarial learning method, which involves two processes in a minimax game: a feature projector that generates modality-invariant and distinguished representations, and a modality classifier that attempts to detect modalities of items given unknown features. It also utilizes triplet constraints to ensure that cross-modal semantic information remains well when projected into a common subspace.
- *CCL* adopts a multi-task learning strategy to adaptively balance the semantic category constraints and the pairwise similarity constraints between both modalities. It also constructs a multi-grained model, which can fuse the coarse-grained cases with fine-grained patches, to tap more information from the inherent intramodality and intermodality correlation.

#### 4.5. Experimental results

This part describes the experimental results and analyses of our HRL method and the comparison methods. Tables 2 and 3 show

the MAP scores of our HRL method and the comparison methods for two retrieval tasks on the Wikipedia dataset using the hand-crafted and CNN feature, respectively. From Tables 2 and 3, the following conclusions can be drawn:

- Our HRL method is superior to the best competitor, CCL, by 74.55% and 36.38% in the average score using the hand-crafted and CNN feature, respectively, for the bi-modal retrieval task. For all-modal retrieval task, HRL performs better than CCL by 48.62% and 28.49% in the average score using the hand-crafted and CNN feature, respectively. This significant performance improvement shows the advantage of relying on hybrid representation learning to tap more complex and richer information about cross-modal correlation.
- Our HRL method is superior to CMDN and CCL, which extract intramodal and intermodal representations to mine the correlation between two modalities, and ACMR, which learns a discriminative and modality-invariant representation. This suggests that learning the modality-friendly representation, which is more similar in terms of its statistical properties than that of the original inputs, and the modality-mutual representation, which involves some missing mutual information in both modalities, is beneficial to improve the performance of cross-modal retrieval.
- Among the comparison methods, the shallow learning method JRL achieves comparable performance with DNN-based methods, and even better than the corr-AE, bimodal AE, and multimodal DBN. This may be because the small size of the Wikipedia dataset is not sufficient for these deep networks to attain an ideal performance.
- For the traditional methods, KCCA has worse performance than CFA in some settings because KCCA can only learn a coarse association between different modalities. In addition, the performance of different kernel functions may vary greatly. In particular, the poly kernel has worse performance than the Gaussian kernel on most settings because it cannot effectively handle the large-scale training data.
- Among the DNN-based methods, multimodal DBN exhibits worst performance among them. It is because only a joint distribution is learned on the top of two-layer DBN for each modality, which focuses more on learning the complementarity between different modalities rather than the correlation across them. The bimodal AE and corr-AE perform better because they jointly consider the reconstruction information.

Tables 4–7 show the retrieval results on the NUS-WIDE-10k dataset and Pascal Sentences dataset. The trends of the retrieval results on these two datasets are similar to the Wikipedia dataset and our HRL method maintains the optimal performance.

Some retrieval results of our HRL method and CCL on the Pascal Sentences dataset are shown in Fig. 5, from which we can see that our proposed HRL method can effectively reduce the failure cases compared with the best competitor CCL. In addition, a few categories in this dataset are difficult to be distinguished due to the similar descriptions such as the bicycle, motorbike and car categories, which lead to confusions during the retrieval process on both our HRL method and the compared approaches. However, our method still performs best with the least failure cases.

In addition to the assessment based on the MAP score, the precision-recall curves are drawn for additional comparisons. In Figs. 6–11, the curves of our HRL method, ACMR, CMDN, JRL, corr-AE, bimodal AE, multimodal DBN, KCCA (Gaussian), KCCA (poly), CFA and CCA using the hand-crafted feature are shown. For the three datasets, except the Wikipedia and NUS-WIDE-10k dataset, the assessments of the precision-recall curves on the Pascal Sentences dataset are inconsistent with the MAP scores for both the image and text retrieval, where our HRL method is better than its

**Table 2**The MAP scores of *bi-modal retrieval* for our HRL method and the comparison methods on the *Wikipedia dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image → Text         | Text → Image | Average      | Image → Text | Text → Image | Average      |
| <b>HRL (proposed)</b> | <b>0.672</b>         | <b>0.686</b> | <b>0.679</b> | <b>0.647</b> | <b>0.666</b> | <b>0.656</b> |
| CCL [41]              | 0.418                | 0.359        | 0.389        | 0.504        | 0.457        | 0.481        |
| ACMR [46]             | 0.315                | 0.237        | 0.276        | 0.429        | 0.374        | 0.401        |
| CMDN [47]             | 0.393                | 0.325        | 0.359        | 0.488        | 0.427        | 0.458        |
| JRL [15]              | 0.344                | 0.277        | 0.311        | 0.453        | 0.400        | 0.427        |
| Corr-AE [29]          | 0.280                | 0.242        | 0.261        | 0.402        | 0.395        | 0.399        |
| Bimodal AE [24]       | 0.236                | 0.208        | 0.222        | 0.314        | 0.290        | 0.302        |
| Multimodal DBN [28]   | 0.149                | 0.150        | 0.150        | 0.204        | 0.183        | 0.194        |
| KCCA (Gaussian) [5]   | 0.245                | 0.219        | 0.232        | 0.326        | 0.268        | 0.297        |
| KCCA (poly) [5]       | 0.200                | 0.185        | 0.193        | 0.215        | 0.214        | 0.215        |
| CFA [10]              | 0.236                | 0.211        | 0.224        | 0.334        | 0.297        | 0.316        |
| CCA [4]               | 0.203                | 0.183        | 0.193        | 0.258        | 0.250        | 0.254        |

**Table 3**The MAP scores of *all-modal retrieval* for our HRL method and the comparison methods on the *Wikipedia dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image → All          | Text → All   | Average      | Image → All  | Text → All   | Average      |
| <b>HRL (proposed)</b> | <b>0.772</b>         | <b>0.628</b> | <b>0.700</b> | <b>0.762</b> | 0.619        | <b>0.690</b> |
| CCL [41]              | 0.331                | 0.610        | 0.471        | 0.422        | <b>0.652</b> | 0.537        |
| ACMR [46]             | 0.266                | 0.604        | 0.435        | 0.381        | 0.546        | 0.463        |
| CMDN [47]             | 0.282                | 0.592        | 0.437        | 0.361        | 0.637        | 0.499        |
| JRL [15]              | 0.281                | 0.556        | 0.419        | 0.381        | 0.530        | 0.456        |
| Corr-AE [29]          | 0.225                | 0.401        | 0.313        | 0.311        | 0.537        | 0.424        |
| Bimodal AE [24]       | 0.175                | 0.422        | 0.299        | 0.281        | 0.517        | 0.399        |
| Multimodal DBN [28]   | 0.140                | 0.177        | 0.159        | 0.170        | 0.190        | 0.180        |
| KCCA (Gaussian) [5]   | 0.163                | 0.377        | 0.270        | 0.321        | 0.472        | 0.397        |
| KCCA (poly) [5]       | 0.158                | 0.317        | 0.238        | 0.256        | 0.320        | 0.288        |
| CFA [10]              | 0.174                | 0.283        | 0.229        | 0.300        | 0.364        | 0.332        |
| CCA [4]               | 0.180                | 0.315        | 0.248        | 0.219        | 0.343        | 0.281        |

**Table 4**The MAP scores of *bi-modal retrieval* for our HRL method and the comparison methods on the *NUS-WIDE-10k dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image→Text           | Text→Image   | Average      | Image→Text   | Text→Image   | Average      |
| <b>HRL (proposed)</b> | <b>0.446</b>         | <b>0.476</b> | <b>0.461</b> | <b>0.603</b> | <b>0.599</b> | <b>0.601</b> |
| CCL [41]              | 0.400                | 0.401        | 0.401        | 0.506        | 0.535        | 0.521        |
| ACMR [46]             | 0.354                | 0.328        | 0.340        | 0.520        | 0.520        | 0.520        |
| CMDN [47]             | 0.391                | 0.357        | 0.374        | 0.492        | 0.515        | 0.504        |
| JRL [15]              | 0.324                | 0.263        | 0.294        | 0.426        | 0.376        | 0.401        |
| Corr-AE [29]          | 0.223                | 0.227        | 0.225        | 0.366        | 0.417        | 0.392        |
| Bimodal AE [24]       | 0.159                | 0.172        | 0.166        | 0.327        | 0.369        | 0.348        |
| Multimodal DBN [28]   | 0.158                | 0.130        | 0.144        | 0.201        | 0.259        | 0.230        |
| KCCA (Gaussian) [5]   | 0.232                | 0.213        | 0.223        | 0.300        | 0.336        | 0.318        |
| KCCA (poly) [5]       | 0.150                | 0.149        | 0.150        | 0.114        | 0.130        | 0.122        |
| CFA [10]              | 0.211                | 0.188        | 0.200        | 0.400        | 0.299        | 0.350        |
| CCA [4]               | 0.141                | 0.138        | 0.140        | 0.202        | 0.220        | 0.211        |

**Table 5**The MAP scores of *all-modal retrieval* for our HRL method and the comparison methods on the *NUS-WIDE-10k dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image → All          | Text → All   | Average      | Image → All  | Text → All   | Average      |
| <b>HRL (proposed)</b> | <b>0.540</b>         | 0.418        | <b>0.479</b> | <b>0.627</b> | <b>0.582</b> | <b>0.605</b> |
| CCL [41]              | 0.379                | 0.444        | 0.412        | 0.537        | 0.502        | 0.520        |
| ACMR [46]             | 0.338                | <b>0.445</b> | 0.392        | 0.564        | 0.531        | 0.548        |
| CMDN [47]             | 0.306                | 0.417        | 0.362        | 0.478        | 0.449        | 0.464        |
| JRL [15]              | 0.237                | 0.421        | 0.329        | 0.445        | 0.357        | 0.401        |
| Corr-AE [29]          | 0.222                | 0.245        | 0.234        | 0.389        | 0.379        | 0.384        |
| Bimodal AE [24]       | 0.145                | 0.257        | 0.201        | 0.255        | 0.287        | 0.271        |
| Multimodal DBN [28]   | 0.128                | 0.171        | 0.150        | 0.193        | 0.338        | 0.266        |
| KCCA (Gaussian) [5]   | 0.147                | 0.282        | 0.215        | 0.386        | 0.351        | 0.369        |
| KCCA (poly) [5]       | 0.138                | 0.173        | 0.156        | 0.304        | 0.150        | 0.227        |
| CFA [10]              | 0.169                | 0.235        | 0.202        | 0.383        | 0.314        | 0.349        |
| CCA [4]               | 0.143                | 0.176        | 0.160        | 0.215        | 0.216        | 0.216        |



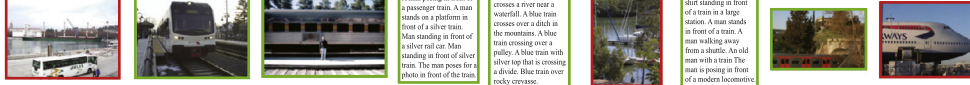





**Table 6**The MAP scores of *bi-modal retrieval* for our HRL method and the comparison methods on the *Pascal Sentences dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image → Text         | Text → Image | Average      | Image → Text | Text → Image | Average      |
| <b>HRL (proposed)</b> | <b>0.389</b>         | <b>0.379</b> | <b>0.384</b> | <b>0.625</b> | <b>0.612</b> | <b>0.619</b> |
| CCL [41]              | 0.359                | 0.346        | 0.353        | 0.566        | 0.560        | 0.563        |
| ACMR [46]             | 0.210                | 0.175        | 0.193        | 0.405        | 0.423        | 0.414        |
| CMDN [47]             | 0.334                | 0.333        | 0.334        | 0.534        | 0.534        | 0.534        |
| JRL [15]              | 0.300                | 0.286        | 0.293        | 0.504        | 0.489        | 0.497        |
| Corr-AE [29]          | 0.268                | 0.273        | 0.271        | 0.489        | 0.484        | 0.487        |
| Bimodal AE [24]       | 0.245                | 0.256        | 0.251        | 0.456        | 0.470        | 0.463        |
| Multimodal DBN [28]   | 0.197                | 0.183        | 0.190        | 0.477        | 0.424        | 0.451        |
| KCCA (Gaussian) [5]   | 0.233                | 0.249        | 0.241        | 0.361        | 0.325        | 0.343        |
| KCCA (poly) [5]       | 0.207                | 0.191        | 0.199        | 0.209        | 0.192        | 0.201        |
| CFA [10]              | 0.187                | 0.216        | 0.202        | 0.351        | 0.340        | 0.346        |
| CCA [4]               | 0.105                | 0.104        | 0.105        | 0.169        | 0.151        | 0.160        |

**Table 7**The MAP scores of *all-modal retrieval* for our HRL method and the comparison methods on the *Pascal Sentences dataset*.

| Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                       | Image → All          | Text → All   | Average      | Image → All  | Text → All   | Average      |
| <b>HRL (proposed)</b> | <b>0.610</b>         | 0.404        | <b>0.507</b> | <b>0.702</b> | <b>0.623</b> | <b>0.662</b> |
| CCL [41]              | 0.352                | <b>0.516</b> | 0.434        | 0.554        | 0.615        | 0.585        |
| ACMR [46]             | 0.334                | 0.453        | 0.393        | 0.426        | 0.528        | 0.477        |
| CMDN [47]             | 0.328                | 0.497        | 0.413        | 0.532        | 0.604        | 0.568        |
| JRL [15]              | 0.316                | 0.459        | 0.388        | 0.501        | 0.563        | 0.532        |
| Corr-AE [29]          | 0.305                | 0.367        | 0.336        | 0.475        | 0.558        | 0.517        |
| Bimodal AE [24]       | 0.263                | 0.417        | 0.340        | 0.466        | 0.558        | 0.512        |
| Multimodal DBN [28]   | 0.208                | 0.323        | 0.266        | 0.459        | 0.413        | 0.436        |
| KCCA (Gaussian) [5]   | 0.224                | 0.416        | 0.320        | 0.423        | 0.540        | 0.482        |
| KCCA (poly) [5]       | 0.218                | 0.446        | 0.332        | 0.335        | 0.261        | 0.298        |
| CFA [10]              | 0.206                | 0.395        | 0.301        | 0.384        | 0.427        | 0.406        |
| CCA [4]               | 0.196                | 0.226        | 0.211        | 0.334        | 0.232        | 0.283        |

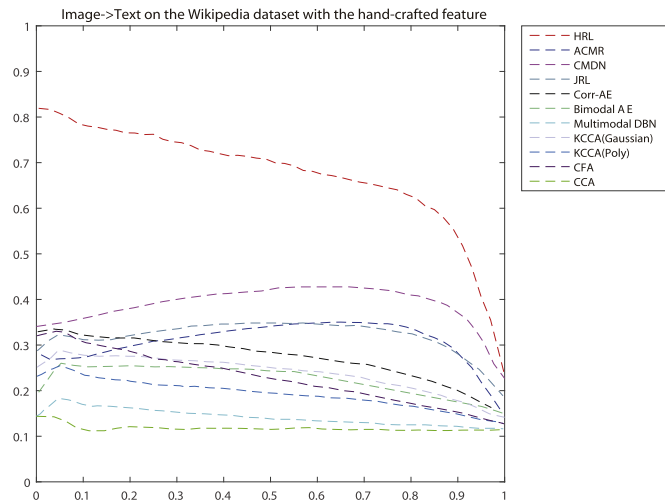
| Query  |     | Results  |  |  |  |  |  |  |  |  |  |
|--|-----|--|--|--|--|--|--|--|--|--|--|
| <br>train   | HRL |  |  |  |  |  |  |  |  |  |  |
|  | CCL |  |  |  |  |  |  |  |  |  |  |
| <br>bicycle | HRL |  |  |  |  |  |  |  |  |  |  |
|  | CCL |  |  |  |  |  |  |  |  |  |  |

**Fig. 5.** Examples of the *all-modal retrieval* results with the hand-crafted feature on the *Pascal Sentences dataset* by our HRL method and CCL. It should be noted that, in these examples, the correct results are shown with green borders, while the wrong results have red borders.

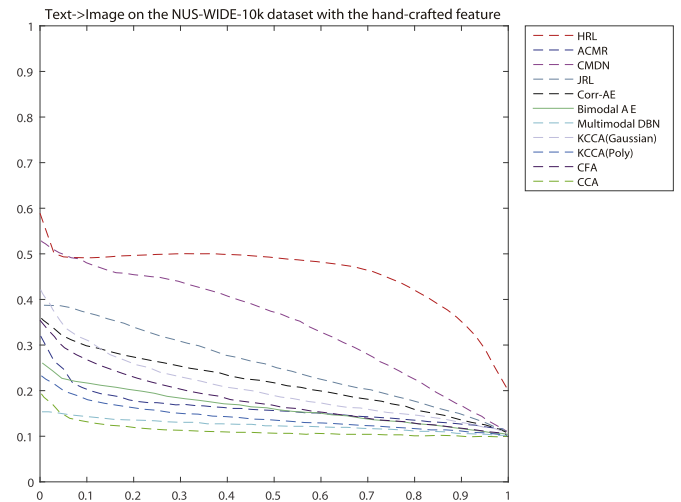
counterparts. The reason for the relatively poor curves on the *Pascal Sentences dataset* is due to the small size of the dataset.

The above results show the stable advantage of our HRL compared with existing methods. It should be noted that our proposed HRL method performs best in the following 3 aspects:

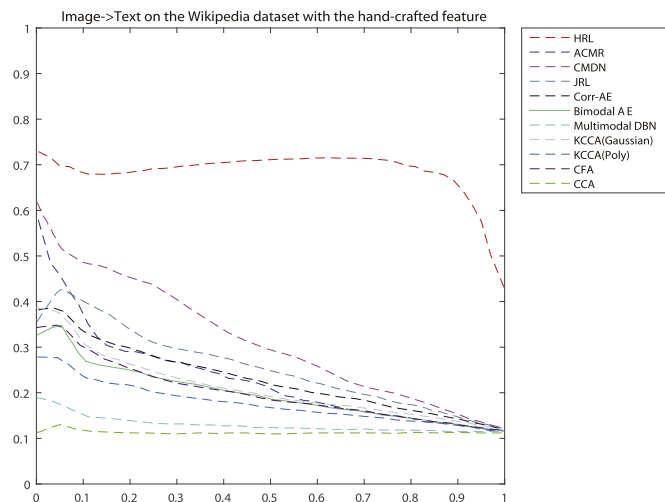
- (1) The statistical properties of modality-friendly representation are more similar than the intramodality representation proposed by the compared methods.
- (2) The modality-mutual representation contains some missing information in the original input modality, while the compared methods only use the original modality in-



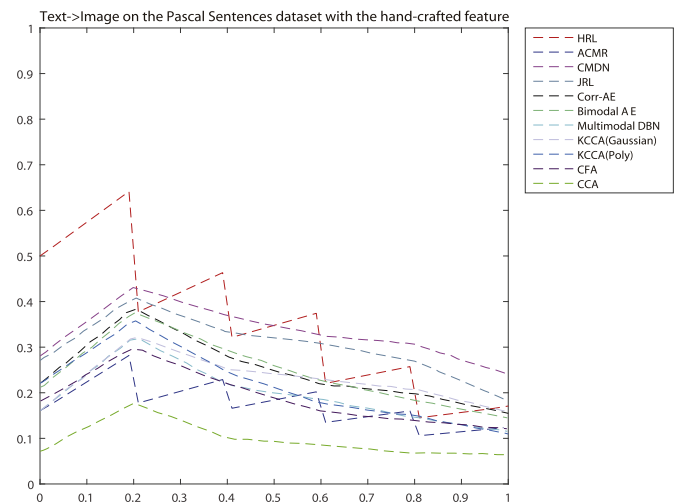
**Fig. 6.** The precision-recall curves of Image  $\rightarrow$  Text retrieval on the Wikipedia dataset.



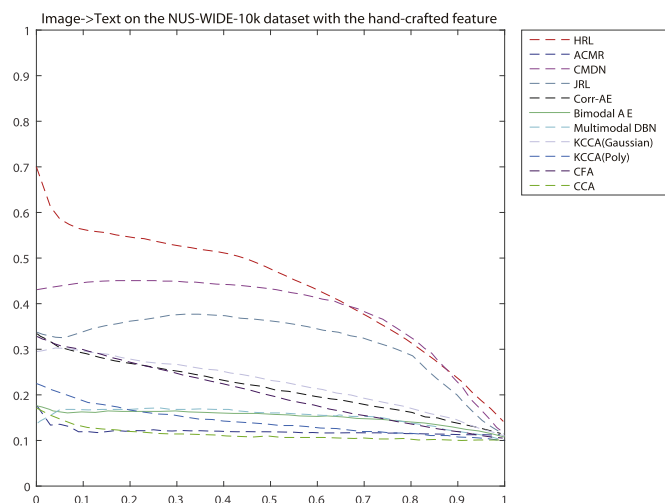
**Fig. 9.** The precision-recall curves of Text  $\rightarrow$  Image retrieval on the NUS-WIDE-10k dataset.



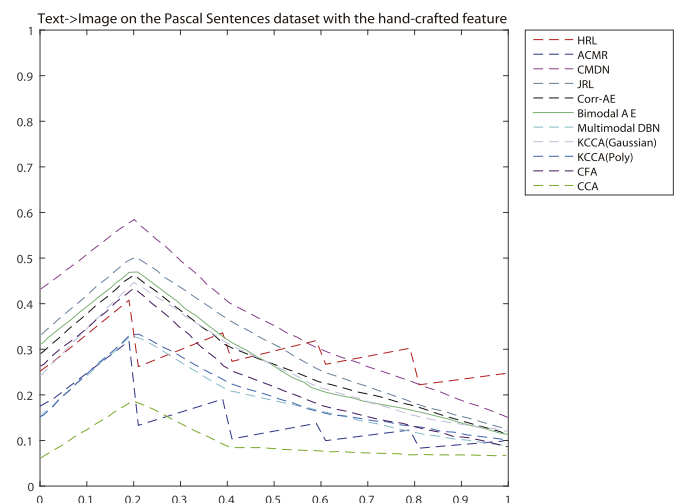
**Fig. 7.** The precision-recall curves of Text  $\rightarrow$  Image retrieval on the Wikipedia dataset.



**Fig. 10.** The precision-recall curves of Image  $\rightarrow$  Text retrieval on the Pascal Sentences dataset.



**Fig. 8.** The precision-recall curves of Image  $\rightarrow$  Text retrieval on the NUS-WIDE-10k dataset.



**Fig. 11.** The precision-recall curves of Text  $\rightarrow$  Image retrieval on the Pascal Sentences dataset.

**Table 8**The MAP scores of *bi-modal retrieval* for joint-HRL and our HRL method on three datasets.

| Dataset          | Methods               | hand-crafted feature |              |              | CNN feature  |              |              |
|------------------|-----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|                  |                       | Image → Text         | Text → Image | Average      | Image → Text | Text → Image | Average      |
| Wikipedia        | Joint-HRL             | 0.378                | 0.285        | 0.331        | 0.413        | 0.357        | 0.385        |
|                  | <b>HRL (proposed)</b> | <b>0.672</b>         | <b>0.686</b> | <b>0.679</b> | <b>0.647</b> | <b>0.666</b> | <b>0.656</b> |
| NUS-WIDE-10k     | Joint-HRL             | 0.418                | 0.393        | 0.405        | 0.522        | 0.536        | 0.529        |
|                  | <b>HRL (proposed)</b> | <b>0.446</b>         | <b>0.476</b> | <b>0.461</b> | <b>0.603</b> | <b>0.599</b> | <b>0.601</b> |
| Pascal Sentences | Joint-HRL             | 0.227                | 0.191        | 0.209        | 0.570        | 0.576        | 0.573        |
|                  | <b>HRL (proposed)</b> | <b>0.389</b>         | <b>0.379</b> | <b>0.384</b> | <b>0.625</b> | <b>0.612</b> | <b>0.619</b> |

**Table 9**

Baseline experiments on *bi-modal retrieval*, where HRL (no stacked BAEs) means HRL does not use stacked BAEs, HRL (ff) means HRL does not use feedforward network and HRL (no ff & stacked BAEs) means HRL does not use both feedforward network and stacked BAEs. In addition, HRL (no SRBM & joint-ae) means HRL uses only modality-mutual representation in the first step, while HRL (no MDBN & joint-ae) means HRL uses only modality-friendly representation in the first step. Both these two baseline experiments do not use joint autoencoder in the second step.

| Dataset      | Method                     | hand-crafted feature |              |              | CNN feature  |              |              |
|--------------|----------------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|              |                            | Image → Text         | Text → Image | Average      | Image → Text | Text → Image | Average      |
| Wikipedia    | <b>HRL (proposed)</b>      | <b>0.672</b>         | <b>0.686</b> | <b>0.679</b> | <b>0.647</b> | <b>0.666</b> | <b>0.656</b> |
|              | HRL (no stacked BAEs)      | 0.661                | 0.672        | 0.667        | 0.629        | 0.643        | 0.636        |
|              | HRL (no ff)                | 0.404                | 0.407        | 0.405        | 0.380        | 0.387        | 0.384        |
|              | HRL (no ff & stacked BAEs) | 0.183                | 0.190        | 0.187        | 0.175        | 0.179        | 0.177        |
|              | HRL (no SRBM & joint-ae)   | 0.364                | 0.279        | 0.321        | 0.430        | 0.373        | 0.401        |
|              | HRL (no MDBN & joint-ae)   | 0.338                | 0.275        | 0.307        | 0.471        | 0.420        | 0.445        |
| NUS-WIDE-10k | <b>HRL (proposed)</b>      | <b>0.446</b>         | <b>0.476</b> | <b>0.461</b> | <b>0.603</b> | <b>0.599</b> | <b>0.601</b> |
|              | HRL (no stacked BAEs)      | 0.404                | 0.407        | 0.405        | 0.555        | 0.550        | 0.553        |
|              | HRL (no ff)                | 0.320                | 0.330        | 0.325        | 0.123        | 0.146        | 0.138        |
|              | HRL (no ff & stacked BAEs) | 0.122                | 0.123        | 0.122        | 0.125        | 0.125        | 0.125        |
|              | HRL (no SRBM & joint-ae)   | 0.418                | 0.397        | 0.408        | 0.526        | 0.547        | 0.537        |
|              | HRL (no MDBN & joint-ae)   | 0.179                | 0.190        | 0.184        | 0.217        | 0.251        | 0.234        |

stances. (3) Jointly combining the modality-friendly and modality-mutual representations can not only consider the correlation between the intermodality and intramodality representation proposed by the compared methods but can also fully take advantage of some missing information to enrich the input information.

#### 4.6. Baseline experiments

For further analyses on HRL, Tables 8 and 9 show the results of our HRL method and the baseline methods. Table 8 shows the MAP scores of the variant of HRL and our HRL method. Joint-HRL means that HRL does not generate the hybrid representation, which can use some missing information in the original modality instances to enrich the input information. In other words, joint-HRL generates the joint representation, which combines the image representation constructed by the image-pathway SRBM and the modality-mutual representation which involves obvious image representation, or vice-versa. From Table 8, it should be noted that our HRL method is better than joint-HRL, which is superior to most of the comparison methods. This suggests that utilizing the missing information in the original input instances can enrich the input instances and improve the performance of the retrieval results.

In addition, it should be noted that in Table 8, for the Wikipedia dataset, the performance of HRL with the hand-crafted feature is superior to that of HRL with the CNN feature. This likely because the input of the text modality for the Wikipedia dataset is an article while that of the NUS-WIDE-10k and Pascal Sentences dataset is tags or sentences, respectively. During the test phase, more input text information may play a more important role in the result of the Text → Image retrieval task in the experimental setting with the CNN feature. In addition, during the training phase, more input text information can enrich the shared representation of the image, especially by utilizing missing image information generated by the input text modality. More input text information can not only compensate for the disadvantage of hand-crafted feature, which does

not pay more attention to the category information compared with the CNN feature, but can also offer more missing useful image information for the hand-crafted feature than the CNN feature.

Some baseline experiments are also conducted to verify the effectiveness of the combination of the three steps used by our HRL method. In Table 9, HRL (no stacked BAEs) means HRL does not use stacked bimodal AEs containing label information to obtain the final shared representation, while HRL (no ff) means HRL does not utilize a feedforward network to further obtain the hybrid representation with label information. In addition, HRL (no ff & stacked BAEs) means HRL does not use a feedforward network and stacked BAEs. We can see that the lack of label information in these three baseline experiments makes HRL perform worse. Moreover, the total lack of label information in HRL (no ff & stacked BAEs) makes HRL perform the worst. It is because label information can optimize our HRL method by supervised learning, which can project various modalities into a common representation space.

In addition, HRL (no SRBM & joint-ae) means HRL with only the modality-mutual representation obtained in the first step, while HRL (no MDBN & joint-ae) means HRL with only the modality-friendly representation obtained in the first step. In addition, these two baseline experiments do not use the joint autoencoder in the second step. We can see that learning the modality-friendly and modality-mutual representation correlation simultaneously achieves a higher MAP score than learning only one of them, which indicates that the complementarity of the two kinds of information can be effectively exploited by our HRL method to perform better.

The above baseline experimental results verify the contribution of each step in our HRL method. First, utilizing the missing information in the original input instances can enrich the input instances and improve the performance of the retrieval results. Second, fusion of complementary modality-friendly and modality-mutual representations can lead to a more useful cross-modal shared representation. Third, label information can provide rich

category information to learn the correlation between different modalities, which can help our HRL method capture the important hints to boost the shared representation.

## 5. Conclusions

In this paper, a new method (HRL) is proposed to learn the cross-modal shared representation. In the first learning step, HRL learns the modality-friendly representation, whose statistical properties are similar, and the modality-mutual representation, which contains some missing information in the original input instances. In the second learning step, HRL utilizes a two-level network containing a joint autoencoder and a three-layer feedforward neural net to obtain the hybrid representation, which not only considers the correlation between the intermodality and intramodality representations but also uses some missing information to enrich the input information. In the third learning step, HRL obtains the final shared representation for each modality by implementing stacked bimodal autoencoders. The experimental results show that our proposed HRL method is superior to eleven state-of-the-art approaches according to three widely used cross-modal datasets.

There are two aspects to our future work. First, some semi-supervised strategies will be applied in our method, which may be beneficial to obtain a better high-level shared representation and to further learn the rich and complex correlation between various modalities. Second, we will continue to concentrate on utilizing and combining other deep neural networks to achieve better performance on the cross-modal retrieval task because various representations can be acquired by various networks.

## Acknowledgments

This work was supported in part by the [National Natural Science Foundation of China](#) under Grant 61771322 and 61375015, and the Shenzhen Foundation fund No.JCYJ20160307154630057.

## References

- [1] C. Deng, X. Tang, J. Yan, W. Liu, X. Gao, Discriminative dictionary learning with common label alignment for cross-modal retrieval, *IEEE Trans. Multimed.* 18 (2) (2016) 208–218.
- [2] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, 2018. doi:10.1109/TPAMI.2018.2852750.
- [3] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, *IEEE Trans. Multimed.* 17 (11) (2015) 1989–1999.
- [4] H. Hotelling, Relations between two sets of variates, *Biometrika* 28 (3/4) (1936) 321–377.
- [5] D.R. Hardoon, S. Szedmak, J. Shawetaylor, Canonical correlation analysis: An overview with application to learning methods, *Neural Comput.* 16 (12) (2004) 2639–2664.
- [6] H. Bredin, G. Chollet, Audio-visual speech synchrony measure for talking-face identity verification, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 233–236.
- [7] B. Klein, G. Lev, G. Sadeh, L. Wolf, Associating neural word embeddings with deep image representations using fisher vectors, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4437–4446.
- [8] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 251–260.
- [9] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, *Int. J. Comput. Vis.* 106 (2) (2012) 210–233.
- [10] D. Li, N. Dimitrova, M. Li, I.K. Sethi, Multimedia content processing through cross-modal association, in: *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, pp. 604–611.
- [11] V. Ranjan, N. Rasiwasia, C.V. Jawahar, Multi-label cross-modal retrieval, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.
- [12] T.Q.N. Tran, H.L. Borgne, M. Crucianu, Aggregating image and text quantized correlated components, in: *Proceedings of the Computer Vision and Pattern Recognition*, 2016, pp. 2046–2054.
- [13] Y. Hua, S. Wang, S. Liu, Q. Huang, A. Cai, Tina: Cross-modal correlation learning by adaptive hierarchical semantic aggregation, in: *Proceedings of the IEEE International Conference on Data Mining*, 2014, pp. 190–199.
- [14] X. Zhai, Y. Peng, J. Xiao, Heterogeneous metric learning with joint graph regularization for cross-media retrieval, in: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013, pp. 1198–1204.
- [15] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semi supervised regularization, *IEEE Trans. Circuits Syst. Video Technol.* 24 (6) (2014) 965–978.
- [16] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (10) (2016) 2010–2023.
- [17] W. Zhang, K. Liu, W. Zhang, Y. Zhang, J. Gu, Deep neural networks for wireless localization in indoor and outdoor environments, *Neurocomputing* 194 (2016) 279–287.
- [18] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 2873–2879.
- [19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Proceedings of the International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [20] Z. Wu, Y.G. Jiang, X. Wang, H. Ye, X. Xue, Multi-stream multi-class fusion of deep networks for video classification, in: *Proceedings of the ACM on Multimedia Conference*, 2016, pp. 791–800.
- [21] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4242–4251.
- [22] E. Yang, C. Deng, C. Li, W. Liu, J. Li, D. Tao, Shared predictive cross-modal deep quantization, *IEEE Trans. Neural Netw. Learn. Syst.* PP (99) (2018) 1–12.
- [23] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, *IEEE Trans. Image Process.* 27 (8) (2018) 3893–3903.
- [24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: *Proceedings of the International Conference on Machine Learning*, 2011, pp. 689–696.
- [25] J. Kim, J. Nam, I. Gurevych, Learning semantics with deep belief network for cross-language information retrieval, in: *Proceedings of the International Conference on Computational Linguistics*, 2012, pp. 579–588.
- [26] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: *Proceedings of the Neural Information Processing Systems*, 2012, pp. 2222–2230.
- [27] D. Wang, P. Cui, M. Ou, W. Zhu, Deep multimodal hashing with orthogonal regularization, in: *Proceedings of the International Conference on Artificial Intelligence*, 2015, pp. 2291–2297.
- [28] N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, in: *Proceedings of the International Conference on Machine Learning Workshop*, 2012.
- [29] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 7–16.
- [30] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.
- [31] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, 2015, pp. 2267–2273.
- [32] Z. Li, J. Tang, X. He, Robust structured nonnegative matrix factorization for image representation, *IEEE Trans. Neural Netw.* 29 (5) (2018) 1947–1960.
- [33] Z. Li, J. Tang, Unsupervised feature selection via nonnegative spectral analysis and redundancy control, *IEEE Trans. Image Process.* 24 (12) (2015) 5343–5355.
- [34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [35] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, in: *Proceedings of the Neural Information Processing Systems*, 2016, pp. 379–387.
- [36] W. Zhang, X. Yu, X. He, Learning bidirectional temporal cues for video-based person re-identification, 2017. doi:10.1109/TCSVT.2017.2718188.
- [37] W. Zhang, B. Ma, K. Liu, R. Huang, Video-based pedestrian re-identification by adaptive spatio-temporal appearance model, *IEEE Trans. Image Process.* 26 (4) (2017) 2042–2054.
- [38] P. Smolensky, *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, MIT Press, 1986.
- [39] M. Welling, M. Rosenzvi, G.E. Hinton, Exponential family harmoniums with an application to information retrieval, in: *Proceedings of the Neural Information Processing Systems*, 2005, pp. 1481–1488.
- [40] G.E. Hinton, R. Salakhutdinov, Replicated softmax: an undirected topic model, in: *Proceedings of the Neural Information Processing Systems*, 2009, pp. 1607–1614.
- [41] Y. Peng, J. Qi, X. Huang, Y. Yuan, CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network, *IEEE Trans. Multimed.* 20 (2) (2018) 405–420.
- [42] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, 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*, 2009, p. 48.
- [43] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 139–147.
- [44] W. Zhang, Q. Chen, W. Zhang, X. He, Long-range terrain perception using convolutional neural networks, *Neurocomputing* 275 (2018) 781–787.



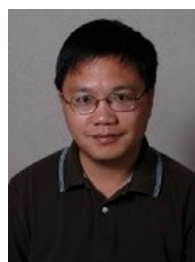
- [45] M. Simon, E. Rodner, J. Denzler, ImageNet pre-trained models with batch normalization, CoRR (2016). <http://arxiv.org/abs/1612.01452>.
- [46] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the ACM on Multimedia Conference, 2017, pp. 154–162.
- [47] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: Proceedings of the International Joint Conferences Artificial Intelligence, 2016, pp. 4846–4853.



**Wenming Cao** received the M.S. degree from the System Science Institute, China Science Academy, Beijing, China, in 1991, and the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 2003. From 2005 to 2007, he was a Post-Doctoral Researcher with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with Shenzhen University, Shenzhen, China. He has authored or coauthored over 80 publications in top-tier conferences and journals. His research interests include pattern recognition, image processing, and visual tracking.



**Qiubin Lin** received the B.S. in Electronic Information Engineering from Shenzhen University, China, in 2017. He is currently pursuing the Master Engineering degree in information and communication engineering, Shenzhen University, Shenzhen, China. His current research interests include deep learning and multimodal retrieval.



**Zhihai He** is a professor in the Electrical Engineering and Computer Science Department at the University of Missouri. He worked as a research engineer at the David Sarnoff Research Center before joining the MU faculty. He was named Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2015 for contributions to video communication and visual sensing technologies.



**Zhiquan He** is currently working as assistant professor in College of Information Engineering, Shenzhen University, China. He received his M.S. degree from Institute of Electronics, Chinese Academy of Sciences in 2001, and the Ph.D degree from the department of Computer Science, University of Missouri-Columbia in 2014. His research area is in the areas of image processing, computer vision and machine learning.