



# Multi-modal semantic autoencoder for cross-modal retrieval

Yiling Wu<sup>a,b</sup>, Shuhui Wang<sup>a,\*</sup>, Qingming Huang<sup>b</sup>

<sup>a</sup>Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

<sup>b</sup>School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China



## ARTICLE INFO

### Article history:

Received 23 June 2018

Revised 28 September 2018

Accepted 12 November 2018

Available online 20 November 2018

Communicated by Yongdong Zhang

### Keywords:

Cross-modal retrieval

Multi-modal data

Autoencoder

## ABSTRACT

Cross-modal retrieval has gained much attention in recent years. As the research mainstream, most of existing approaches learn projections for data from different modalities into a common space where data can be compared directly. However, they neglect the preservation of feature and semantic information, so they are unable to obtain satisfactory results as expected. In this paper, we propose a two-stage learning method to learn multi-modal mappings that project multi-modal data to low dimensional embeddings that preserve both feature and semantic information. In the first stage, we combine both low-level feature and high-level semantic information to learn feature-aware semantic code vectors. In the second stage, we use encoder–decoder paradigm to learn projections. The encoder projects feature vectors to code vectors, and the decoder projects code vectors back to feature vectors. The encoder-decoder paradigm guarantees the embeddings to preserve both feature and semantic information. An alternating minimization procedure is developed to solve the multi-modal semantic autoencoder optimization problem. Extensive experiments on three benchmark datasets demonstrate that the proposed method outperforms state-of-the-art cross-modal retrieval methods.

© 2018 Elsevier B.V. All rights reserved.

## 1. Introduction

In recent years, as a series of progresses have been made in multimedia community [1], multimedia data, such as images, texts, are proliferated on the Web. Faced with massive multimedia data, users often expect to get content in various modalities that fits to their demand. For instance, a user may want to get images depicting the content of a text or texts describing the content of an image. However, traditional retrieval system [2] can not solve the problem, since they can only return documents in the same modality as the queries. Cross-modal retrieval [3–8] has since emerged as a promising paradigm towards the above problem. Unlike traditional retrieval techniques, cross-modal retrieval utilizes the widely available data from various modalities and meets the users' demands in receiving documents in different modalities from queries.

To perform cross-modal retrieval, the key problem is to measure the semantic similarity between data in different modalities. The heterogeneity among different modalities makes the similarity measurement problem to be non-trivial. To solve the problem, the classical method is to learn mapping functions projecting data into a common space [3–6]. In the common space, data from different

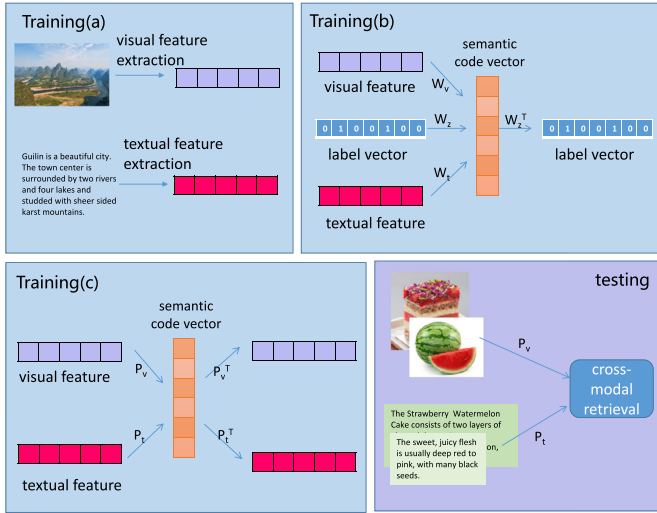
modalities have the same dimensionality, thus simple metric can be used to calculate the similarity.

To achieve favorable retrieval results, embeddings in the common space should preserve both semantic and original feature information. Semantic information [9,10] is the most important information embeddings should preserve as the documents are returned based on semantic similarity to the queries. Moreover, although data in different modalities have different feature spaces, they share the same semantic space. Data with the same semantics are associated whatever the modalities they are from, thus, semantic information can be used to represent not only inter-modality relation, but also intra-modality relation. Original feature information [11–13] complements semantic information by providing additional intra-modality relation. Keeping original feature information exploits intrinsic geometry of each modality and makes the projections to be smoother [14,15]. Thus, good embeddings should preserve semantic information and original feature information simultaneously.

Some existing cross-modal retrieval methods concern the relations of data embeddings in the common space, such as maximizing statistical correlation [16–18], minimizing distance [5,19], but do not pay attention to the information that embeddings contain. Other cross-modal retrieval methods which consider preserving information only preserve semantic information or original feature information. Corr-AE [12] which uses autoencoder to reconstruct both modalities focuses on preserving original feature information.

\* Corresponding author.

E-mail addresses: [yiling.wu@vipl.ict.ac.cn](mailto:yiling.wu@vipl.ict.ac.cn) (Y. Wu), [wangshuhui@ict.ac.cn](mailto:wangshuhui@ict.ac.cn) (S. Wang), [qmhuang@ucas.ac.cn](mailto:qmhuang@ucas.ac.cn) (Q. Huang).



**Fig. 1.** Framework of the proposed method. Left figure shows the training phase, and right figure shows the testing phase. In the training phase, we first learn feature-aware semantic code vector which contains both feature information and label information. Then we learn projections by multi-modal semantic autoencoder which jointly projects image and text to the learned code vector and reconstructs image and text from code vector.

The lack of semantic information leads to limited retrieval results. SCM [9], LCFS [10] and LGCF [20] which perform regression from feature spaces to label-based space focus on preserving semantic information. However, SCM, LCFS and LGCF deal with single-class situation, therefore they ignore the correlation between labels in multi-label settings. Moreover, they fix the dimension of common space to the label space, which results in inefficiency when the dimensionality of label space is large.

To address the above problems, we propose a method to learn projections which project data to embeddings preserving both original feature information and semantic information. Without loss of generality, we use image and text modalities for illustration. The information preservation is ensured by two steps. At the first step, we extend conditional principal label space transformation (CPLST) [21] to multi-modal situation to learn feature-aware semantic code vector. The code vector contains both feature and semantic information, and is used to direct the learning of projections in the second stage. Different from LCFS and LGCF which fix the dimension of common space to label space, our method compresses the label space providing more flexible choice of common space dimension, and drops the redundant information. At the second step, we take the encoder-decoder paradigm to learn projections. There are two pairs of encoder and decoder, i.e., one pair for image modality and one pair for text modality. The encoders project images and texts to the semantic code vectors. The decoders exert an additional constraint, that is, the code must be able to reconstruct both visual and textual original features. Thus the encoders are taken as projections projecting features to embeddings preserving both semantic information and original feature information. The encoders and decoders are linear and symmetric with respect to its counterpart which enable us to develop an efficient learning algorithm. The two steps of our method are related. The uniformed code vectors learned in the first step take in visual, textual and semantic information, thus facilitate the learning of projections in the second step. Fig. 1 shows the framework of our method.

The rest of this paper is organized as follows. We discuss related work in Section 2. We describe our proposed approach in Section 3. We demonstrate the superior performance of our method in Section 4. Finally, we conclude our method in Section 5.

## 2. Related work

### 2.1. Cross-modal retrieval

Due to the widespread existence of multi-modal data, cross-modal retrieval has drawn much attention in recent years. Various approaches have been proposed to deal with cross-modal retrieval task. Generally speaking, since modalities are heterogeneous, the largest difference of cross-modal learning methods from single-modal learning methods is that multiple mapping functions should be learned. Each modality has a modality-specific mapping function to project data into the common space. Cross-modal learning methods usually simultaneously utilize many techniques, so they should be classified differently according to different classification rules. The cross-modal learning is also treated as modality-gap bridging and knowledge transfer process, which is beneficial for more comprehensive image understanding [22].

Based on the representation in common space, two categories of the cross-modal learning approaches can be distinguished, real-valued representation learning methods [4,9,23–25] and binary representation learning methods [5,26–29]. Distributed hashing approach [30] and metric learning [31] have also been proposed for large-scale multimedia search that involves multiple modalities, and promising performance has been achieved on benchmark datasets. Most of the cross-modal learning approaches concentrating on learning high quality representation fall into the first category. Some other methods pay attention to storage costs and retrieval efficiency learning binary representation which is usually called hash code. Our method belongs to real-valued representation learning method.

Inspired by the success of deep representation learning, a lot of methods using deep learning have been proposed recently. According to whether to use neural network, cross-modal correlation learning methods can be divided into two categories, traditional methods and deep methods. Traditional methods use linear mapping functions [4,9,23,24,29] and extend to non-linear case by adopting kernel trick [3,4]. Deep methods use neural network as mapping functions in order to use the great representation power of neural network [19,32–34]. In deep methods, vision part usually contains a deep convolutional neural network (CNN) [19,32,35,36] which generates the image representation. Language model part usually learns a dense feature embedding for each word, and generates the text representation by Word CNN [19,37] or recurrent neural network (RNN) [32,38].

More concrete classification rules are according to the applied techniques. Some cross-modal learning methods employ probabilistic models. Jia et al. [39] propose a probabilistic model by defining a Markov random field over the documents which connects the documents based on their similarity. The topics learned with the model are shared across connected documents. Zhen and Yeung [26] propose multi-modal latent binary embedding (MLBE) to learn hash functions from multi-modal data. It assumes that observed intra-modality and inter-modality similarities are generated from the binary latent factors, intra-modality weighting matrices and inter-modality weighting variable.

Another direction of cross-modal learning research based on canonical correlation analysis (CCA) [9,16]. CCA is a basic method which learns projections by maximizing the correlation between projected data of two modalities. Several methods extend CCA to include semantic information, e.g., GMA [4], ml-CCA [18] and 3view-CCA [17]. GMA [4] combines some unsupervised and supervised feature learning techniques with CCA to produce multi-view feature extraction methods. 3view-CCA [17] extends CCA to incorporate a third view which is semantic information. ml-CCA [18] extends CCA by taking into account the multi-label information and using multi-label information to establish cor-

respondences instead of explicit pairings between two modalities.

Moreover, the learning-to-rank approaches have been widely applied to cross-modal approaches. SSI [40] and PAMIR [41] are proposed to learn similarity functions between a query and a document by taking the framework of pairwise learning-to-rank. Wu et al. [23] propose a method which optimizes a bi-directional listwise ranking loss. It employs the structural SVM to support the optimization of various ranking evaluation measures. Zhang et al. [6] propose to simultaneously optimize pairwise and listwise ranking loss with nuclear norm regularization to obtain a common space.

Besides the above kinds of methods, there are some other methods based on graph-based learning method. Song et al. [27] propose inter-media hashing (IMH) model to explore the correlations among multiple media types from different data sources and tackle the scalability issue. Wang et al. [24] propose to project cross-modal data into label space and impose a multi-modal graph regularization term which preserves the inter-modality and intra-modality similarity relationships.  $\ell_{21}$ -norm penalties are imposed on the projection matrices separately to select relevant and discriminative features.

## 2.2. Autoencoder

Autoencoders [42,43] are models that are trained to learn hidden representations for a set of data. An autoencoder may be viewed as consisting of two parts: an encoder function that produces a hidden representation and a decoder that produces a reconstruction. Autoencoders are trained to copy their inputs to their outputs, so the hidden representations are expected to take on useful properties. Many variants of autoencoders are proposed in the literature. Feng et al. [12] correlate hidden representations of two uni-modal autoencoders. Kodirov et al. [13] restrict the hidden representation to be similar to semantic code vector for zero-shot learning. Yang et al. [44] impose a graph regularized constraint on the hidden representations. The main difference between existing study and this paper is that our semantic autoencoder forces the hidden representation to be similar to semantic code vector, while traditional autoencoder simply seeks representation to reconstruct the original data.

## 3. Proposed method

### 3.1. Problem formulation

Suppose we have a collection of  $n$  image-text pairs. Let  $V = [v_1, v_2, \dots, v_n] \in \mathbb{R}^{d_v \times n}$  and  $T = [t_1, t_2, \dots, t_n] \in \mathbb{R}^{d_t \times n}$  denote the visual and textual feature matrices, respectively, where  $d_v$  and  $d_t$  are the visual and textual feature dimensionalities. Each pair of image and text shares the same semantic content. Let  $Y = [y_1, y_2, \dots, y_n] \in \{0, 1\}^{c \times n}$  denote the label matrix for the  $n$  image-text pairs, where  $c$  is the number of classes or labels. Each pair of image and text may be labeled with exactly one of the  $c$  classes (in single label setting), or alternatively, each pair may be described by several of the  $c$  labels (in multiple label setting).

In this paper, we attempt to learn projection functions to project image and text into a common space where the embeddings preserving semantic information and original feature information simultaneously. To guarantee the projected images and texts containing both semantic information and original feature information, we propose a novel method which involves two steps:

- **Step 1:** Learning feature-aware semantic code vector.
- **Step 2:** Learning projections by multi-modal semantic autoencoder.

Fig. 1 shows the framework of the proposed method. We will introduce the two steps elaborately in the following subsections.

### 3.2. Feature-aware semantic code vector

The challenge of cross-modal retrieval arises in that cross-modal data have significantly different statistical properties. We thus leverage label vector to bridge the gap between image and text by learning code vector in the joint embedding space. We extend conditional principal label space transformation (CPLST) [21] to multi-modal situation to learn feature-aware semantic code vector. CPLST belongs to label space dimension reduction (LSDR) paradigm. It exploits both the label and the feature parts to compress the label space.

We aim to learn code vector in  $d$ -dimensional joint embedding space, where  $d \leq c$ . We first shift each label vector  $y_i$  to  $z_i = y_i - \bar{y}$ ,  $i = 1, \dots, n$ , where  $\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$  is the estimated mean of the label vectors. Let  $Z$  contain  $z_i$  as columns. We minimize prediction error and encoding error simultaneously with the “conditional principal” directions  $W_z \in \mathbb{R}^{d \times c}$ :

$$\min_{W_t, W_v, W_z, W_z^\top = I} (\|W_t T - W_z Z\|_F^2 + \|W_v V - W_z Z\|_F^2 + \|Z - W_z^\top W_z Z\|_F^2), \quad (1)$$

where  $\|\cdot\|_F$  denotes the Frobenius-norm,  $I \in \mathbb{R}^{d \times d}$  is an identity matrix,  $W_t \in \mathbb{R}^{d \times d_t}$  and  $W_v \in \mathbb{R}^{d \times d_v}$ . The first two terms are prediction error terms for image and text respectively, and the last term is encoding error term. Particularly, the prediction error terms consider feature and label information equally. The encoding error term ensures the more important semantic information can be reconstructed. Through our extension, the feature information which is contained in code vector is from both image and text modalities.

After getting  $W_z$ , we take  $W_z$  to linearly map  $Z$  to the code vector  $C$  by  $C = W_z Z$ . The code vector  $C$  is the feature-aware semantic code vector which is used in the next step.

#### 3.2.1. Optimization

For every fixed  $W_z$  in Eq. (1), the optimization problem for  $W_t$  is simply a linear regression from  $T$  to  $W_z Z$ . Then, the optimal  $W_t$  can be computed by a closed-form solution  $W_t = W_z Z T^\dagger$ , where  $T^\dagger$  is the pseudo inverse of  $T$ . When the optimal  $W_t$  is inserted back into Eq. (1), the first term of Eq. (1) becomes:

$$\|W_z Z T^\dagger T - W_z Z\|_F^2 \Rightarrow \text{tr}(W_z Z (I - H_t) Z^\top W_z^\top) \quad (2)$$

where  $H_t = T^\top (T^\top)^\dagger$  and  $X^\dagger$  is the pseudo inverse of  $X$ . Similar steps are applied on  $W_v$ , and let  $H_v = V^\top (V^\top)^\dagger$ .

Then, Eq. (1) becomes:

$$\begin{aligned} \min_{W_z, W_z^\top = I} & \text{tr}(W_z Z (I - H_t) Z^\top W_z^\top + W_z Z (I - H_v) V^\top W_z^\top \\ & - W_z^\top W_z Z Z^\top - Z Z^\top W_z^\top W_z \\ & + W_z^\top W_z Z Z^\top W_z^\top W_z) \Rightarrow \\ & \max_{W_z, W_z^\top = I} \text{tr}(W_z Z (H_t + H_v - I) Z^\top W_z^\top) \end{aligned} \quad (3)$$

The problem in Eq. (3) can be solved by taking the eigenvectors with the largest eigenvalues of  $Z(H_t + H_v - I)Z^\top$  as the rows of  $W_z$ .

### 3.3. Multi-modal semantic autoencoder

Now, let us focus on learning projection matrices  $P_v \in \mathbb{R}^{d \times d_v}$  and  $P_t \in \mathbb{R}^{d \times d_t}$  to project images and texts to embeddings close enough to the code vectors, and at the same time contain enough original feature information.

For text modality, to ensure the hidden representations contain enough information from original textual features, in addition to projecting texts to hidden representations, we expect the hidden

representations to have the ability to recover the original features, i.e.,

$$P_t T = U, T = P_t^T U, \quad (4)$$

where  $U \in \mathbb{R}^{d \times n}$  denotes the representations of the  $n$  training text in  $d$ -dimensional hidden space. This form is an autoencoder which is linear, has tied weight [45] and only one hidden layer. The encoder projects the input feature into the hidden layer with a lower dimension and the decoder projects it back to the original feature space. This additional reconstruction task imposes a new constraint in learning of the projection function so that the projection must preserve all the information contained in the original textual features.

For image modality, we also adopt an autoencoder to let the embeddings contain information from original visual features. Ideally, we hope the representations of image-text pairs in the hidden space to be uniform, because in retrieval phase, when a query is given, documents are sorted according to their similarity to the query. Thus we have

$$P_v V = U, V = P_v^T U. \quad (5)$$

Combining Eqs. (4) and (5), we have a multi-modal autoencoder. Two encoders  $P_v$  and  $P_t$  map images and texts to  $U$ , and two decoders  $P_v^T$  and  $P_t^T$  map  $U$  back to images and texts. Thus, the hidden representations would contain original feature information from both textual and visual modalities.

We relax the constraints and rewrite the objective of multi-modal autoencoder as:

$$\|P_v V - U\|_F^2 + \alpha \|V - P_v^T U\|_F^2 + \|P_t T - U\|_F^2 + \alpha \|T - P_t^T U\|_F^2, \quad (6)$$

where  $\alpha$  is the balance parameter that decides the relative importance of reconstruction.

To ensure the hidden representations contain enough semantic information, we expect the hidden representations to be similar to the code vector learned above, so as to leverage label information to regularize the latent representations of the autoencoder. We use a soft regularizer to enforce latent representations to be similar to code vectors as follows,

$$\|U - C\|_F^2. \quad (7)$$

Combining the above, we have the multi-modal semantic autoencoder minimization problem in the following form:

$$\min_{P_v, P_t, U} \|P_v V - U\|_F^2 + \alpha \|V - P_v^T U\|_F^2 + \|P_t T - U\|_F^2 + \alpha \|T - P_t^T U\|_F^2 + \beta \|U - C\|_F^2, \quad (8)$$

where  $\beta$  is the balance parameter that decides the relative importance of keeping to code vector.

### 3.3.1. Optimization

Since the objective function in Eq. (8) is not jointly convex with respect to  $P_v$ ,  $P_t$  and  $U$ , in this subsection, we derive an iterative algorithm to update each variable when fixing others alternatively.

**Optimization for  $P_v$  and  $P_t$ .** It is apparent from Eq. (8) that the positions of  $P_v$ ,  $P_t$  are very similar. Thus, in the following we only show the optimization of  $P_v$  as an example. Take the derivative of Eq. (8) with respect to  $P_v$  and set it to zero. We obtain

$$(P_v V - U)V^T + \alpha U(U^T P_v - V^T) = 0,$$

which can be rewritten as

$$P_v V V^T + \alpha U U^T P_v - (1 + \alpha) U V^T = 0. \quad (9)$$

It is a well-known Sylvester equation which can be solved efficiently by the Bartels–Stewart algorithm [13,46].

Similarly, we obtain the equation for  $P_t$

$$P_t T T^T + \alpha U U^T P_t - (1 + \alpha) U T^T = 0. \quad (10)$$

**Optimization for  $U$ .** Take the derivative of Eq. (8) with respect to  $U$  and set it to zero. We obtain

$$(U - P_v V) + \alpha P_v (P_v^T U - V) + (U - P_t T) + \alpha P_t (P_t^T U - T) + \beta (U - C) = 0.$$

We reorganise the above equation, and get

$$U = (\alpha P_v P_v^T + \alpha P_t P_t^T + (2 + \beta)I)^{-1} \cdot ((1 + \alpha)P_v V + (1 + \alpha)P_t T + \beta C). \quad (11)$$

Algorithm 1 summarizes the alternating minimization procedure to optimize Eq. (8).

---

#### Algorithm 1: Multi-modal semantic autoencoder.

---

**Input:** Feature matrices  $V$ ,  $T$  and code vector matrix  $C$ , trade-off parameters  $\alpha$  and  $\beta$

**Output:** Projection matrices  $P_v$  and  $P_t$

Initialize  $U$  by  $C$ ;

**repeat**

    Fix  $U$ , update  $P_v$  by solving Eq.(9);

    Fix  $U$ , update  $P_t$  by solving Eq.(10);

    Fix  $P_v$  and  $P_t$ , update  $U$  by Eq.(11);

**until** convergence;

---

### 3.4. Complexity analysis

Let  $d_m = \max\{d_v, d_t, d\}$ , and assume  $n \geq d$ . For the first stage, the complexity of eigenvalue decomposition is  $\mathcal{O}(n^3)$ . When  $n$  is large, iterative algorithms, such as power iteration, can be used to calculate the eigenpair with largest eigenvalue. In our case, we need eigenvectors with the largest  $d$  eigenvalues, while other eigenvectors can be obtained using deflation. For the second stage, the complexity of Eq. (9) or Eq. (10) depends on the size of feature dimension ( $\mathcal{O}(d_v^3)$  or  $\mathcal{O}(d_t^3)$ ), and the complexity of Eq. (11) is  $\mathcal{O}(nd_m^2)$ . So, the complexity of the second stage is  $\mathcal{O}(knd_m^2)$ , where  $k$  is the number of iterations.

## 4. Experiments

In this section, we provide comprehensive experimental results to show the performance of our method on three widely-used multi-modal datasets. We first compare our proposed method with different baseline methods to verify its effectiveness. Then we conduct additional evaluations to investigate the proposed method in more detail. We denote our method by MMSAE (multi-modal semantic autoencoder). Codes have been made publicly available at <https://github.com/yiling2018/mmsae>.

### 4.1. Datasets

In this subsection, we briefly introduce the datasets we adopted to evaluate our method.

**WIKI [9].** This dataset was collected from “Wikipedia featured articles”. It contains 2866 image-text pairs belonging to 10 semantic categories. As in [9], we use 2173 image-text pairs for training and 693 image-text pairs for testing. The texts are represented by 10-dim LDA features [9,47], and the images are represented by the 1001-dim CNN features extracted from ‘Logits’ layer of Inception-v4 [48] using tensorflow [49].



**Table 1**

The performance comparison in terms of MAP@R on WIKI dataset.

Method	Task					
	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.465	0.551	0.508	0.415	0.401	0.408
PLS	0.445	0.549	0.497	0.417	0.410	0.413
GMLDA	0.458	0.549	0.504	0.414	0.401	0.407
3view-CCA	0.439	0.521	0.480	0.372	0.354	0.363
ml-CCA	0.467	0.537	0.502	0.430	0.403	0.417
LGCFL	0.473	0.548	0.510	0.442	0.418	0.430
JFSSL	0.471	0.539	0.505	0.443	0.414	0.429
MMSAE	<b>0.488</b>	<b>0.581</b>	<b>0.535</b>	<b>0.469</b>	<b>0.442</b>	<b>0.455</b>

**Table 2**

The performance comparison in terms of MAP@R on MIRFLICKR dataset.

Method	Task					
	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.899	0.861	0.880	0.678	0.678	0.678
PLS	0.896	0.788	0.842	0.662	0.651	0.657
GMLDA	0.895	0.856	0.876	0.668	0.672	0.670
3view-CCA	0.894	0.863	0.879	0.638	0.641	0.640
ml-CCA	0.909	0.869	0.889	0.694	0.698	0.696
LGCFL	0.906	0.858	0.882	0.735	0.739	0.737
JFSSL	0.877	0.831	0.854	0.702	0.710	0.706
MMSAE	<b>0.920</b>	<b>0.884</b>	<b>0.902</b>	<b>0.749</b>	<b>0.753</b>	<b>0.751</b>

**MIRFLICKR** [50]. This dataset originally contains 25,000 instances collected from Flickr, each being an image with its associated textual tags. Each image-text pair is assigned with multiple labels from a total of 38 classes. We use the train-test split provided in the dataset but remove images without textual tags or manually annotated labels. 2000-dim tag frequency features [51] are used for text representations. 1001-dim CNN image features from 'Logits' layer in Inception-v4 are used for image representations.

**NUS-WIDE** [52]. This dataset is a real-world image dataset originally containing 269,648 images with 81 concepts. Each image is associated with user tags, which can be seen as an image-text pair. Following [53], we select the top 10 most frequent labels and thus 186,577 images are left. In the chosen experimental data, we randomly sample 50,000 image-text pairs for training and 10,000 image-text pairs for testing. We use the publicly available 1000-dim tag occurrence features for text features [52], and extract 1001-dim CNN from 'Logits' in Inception-v4 for image features.

#### 4.2. Compared methods

We compare our method with seven methods: CCA [16], PLS [54], GMLDA [4], 3view-CCA [17], ml-CCA [18], LGCFL [20] and JFSSL [24].

CCA and PLS are classical methods which use pairwise information to learn a common latent subspace by maximizing the correlation or covariance between projected data of two modalities. GMLDA, 3view-CCA, ml-CCA, LGCFL and JFSSL are supervised methods which exploit the label information. GMLDA combines CCA with LDA. 3view-CCA extends CCA to incorporate a third view which contains label vectors. ml-CCA extends CCA by using label information to establish correspondences between two modalities. LGCFL and JFSSL learn projection matrices to map data into label-based space. LGCFL performs  $\varepsilon$ -dragging on the label space such that the distances between classes can be enlarged, and imposes group sparse constraint in regression process to learn discriminant groups. JFSSL jointly resorts to  $\ell_{21}$ -norm to perform feature selection and imposes graph regularization term on the projected data.

#### 4.3. Evaluation metrics

To evaluate the performance of the proposed method, two directional cross-modal retrieval tasks are conducted: (1) image query versus text database (image-to-text), (2) text query versus image database (text-to-image). The mean average precision (MAP) is adopted to evaluate the effectiveness of different algorithms. In addition, precision-recall curve is adopted to compare the performance of different methods. In the two metrics, if a data point has overlapped labels to a query, the data point is considered to be relevant to the query.

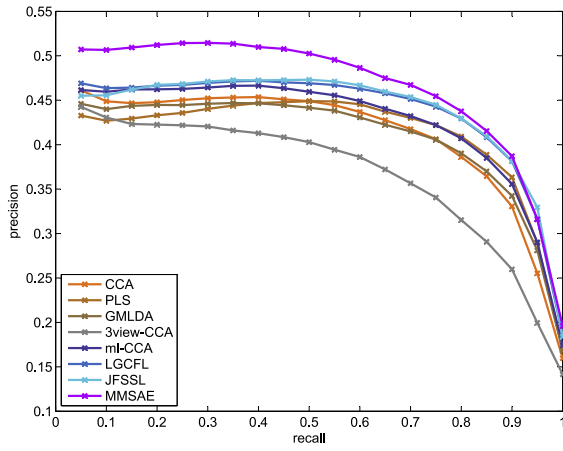
#### 4.4. Parameter setting

To fairly compare with LGCFL and JFSSL, we set the dimensionality of common space to 10, 38 and 10 on WIKI, MIRFLICKR and NUS-WIDE respectively. We use holdout validation to tune  $\alpha$  and  $\beta$  in the range of {10, 1, 0.1, 0.01, 0.001, 0.0001}. For the compared methods, we tune their parameters according to the corresponding literature.

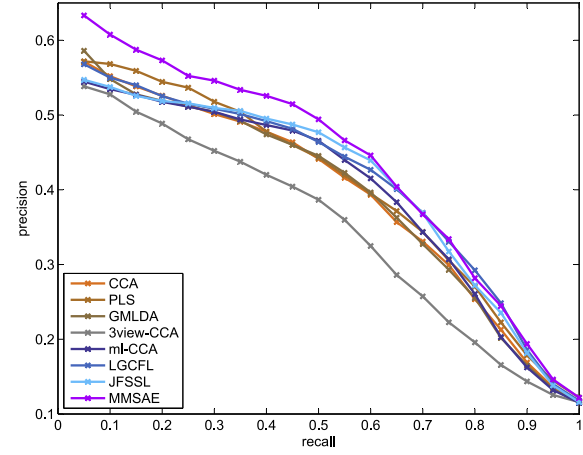
#### 4.5. Comparison with compared methods

Table 1 shows the MAP@R scores achieved by the compared methods and our method on WIKI dataset. We observe that our method MMSAE has remarkable performance gains over the compared methods. This may be because MMSAE learns embeddings preserving original feature and semantic information simultaneously. The semantic information provides both the inter-modality and intra-modality relationships, and the original feature information complementarily provides intra-modality relationship. Fig. 2(a) and (b) show the corresponding precision-recall curves. Again, we observe that compared with its several counterparts, our method obtains better results.

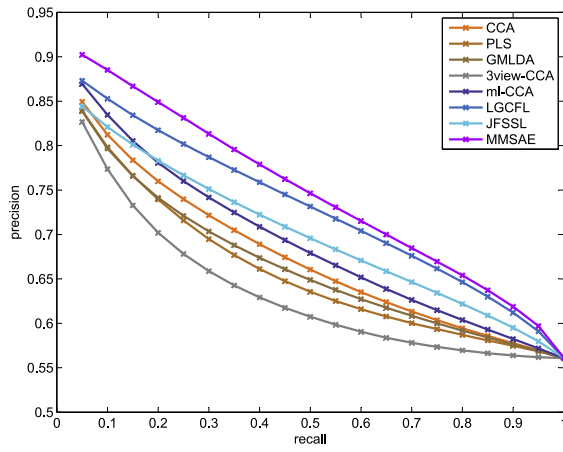
Table 2 shows the MAP@R scores achieved by the compared methods and our method on MIRFLICKR dataset. It is apparent from the table that our method outperforms other compared



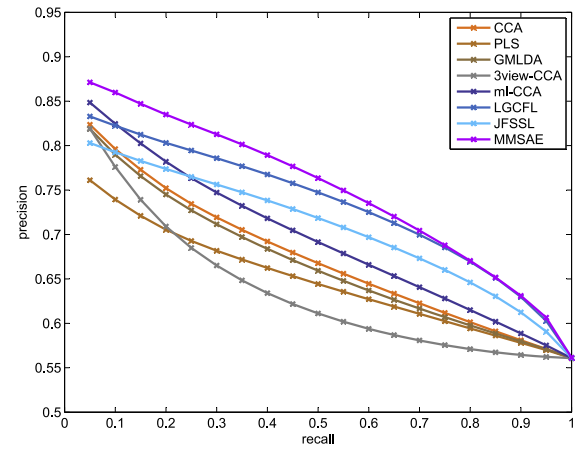
(a) image-to-text on WIKI



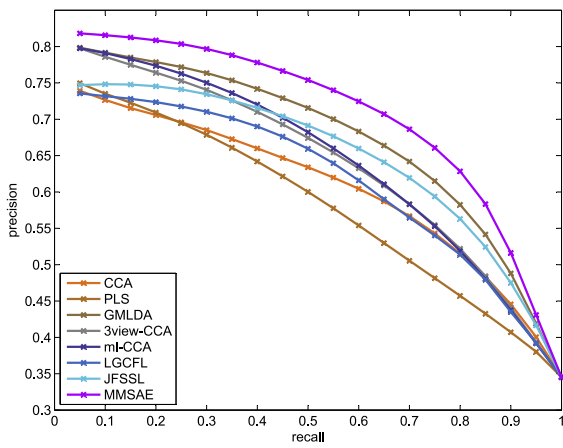
(b) text-to-image on WIKI



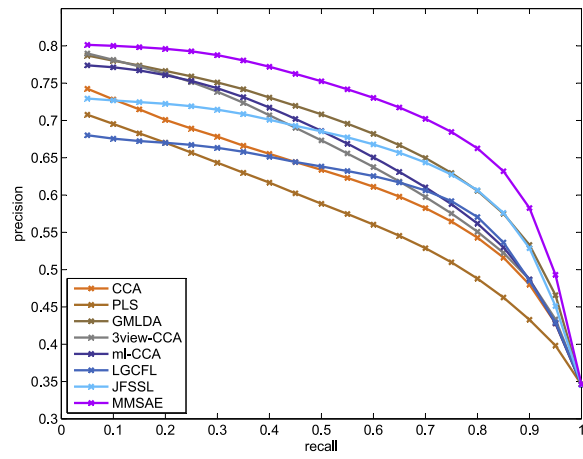
(c) image-to-text on MIRFLICKR



(d) text-to-image on MIRFLICKR



(e) image-to-text on NUS-WIDE




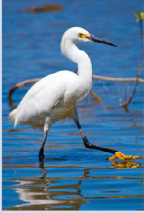
(f) text-to-image on NUS-WIDE

**Fig. 2.** The precision-recall curves of different methods on all benchmark datasets.

**Table 3**

The performance comparison in terms of MAP@R on NUS-WIDE dataset.

Method	Task					
	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.763	0.773	0.768	0.610	0.618	0.614
PLS	0.779	0.736	0.758	0.586	0.554	0.581
GMLDA	0.819	0.810	0.815	0.677	0.680	0.679
3view-CCA	0.824	0.813	0.819	0.644	0.653	0.648
ml-CCA	0.816	0.791	0.804	0.648	0.657	0.652
LGCFL	0.769	0.717	0.743	0.619	0.612	0.616
JFSSL	0.758	0.757	0.758	0.651	0.656	0.653
MMSAE	<b>0.834</b>	<b>0.817</b>	<b>0.826</b>	<b>0.711</b>	<b>0.719</b>	<b>0.715</b>

 abigfave anawesomeshot bokeh goldstaraward naturesfinest impressedbeauty	nature macro flower flowers orange	3
	flower flowers garden iris	3
	flower flowers blueribbonwinner naturesfinest flor flores ltytr1	3
	relevant flower flowers garden purple iris	3
	relevant macro flower	3
 california birds sandiego	water reflection bird birds	5
	nature bird animal naturesfinest birds wildlife bravo avianexcellence flight ave newmexico birdwatcher crane	6
	bird birds	4
	naturesfinest birds avianexcellence specanimal	4
	nature water bird soe outdoors avianexcellence fishing	4

**Fig. 3.** Two examples of image queries and the top 5 texts retrieved by the proposed methods on MIRFLICKR dataset. First column contains the image queries and their corresponded texts. Second column contains retrieved texts. Third column contains numbers of overlapped labels to the query of each retrieved text.

methods. LGCFL achieves the second best result when measured by MAP@all. In our method, we first learn feature-aware semantic code vector from label space. In LGCFL,  $\varepsilon$ -dragging is performed on the label space to force the regression targets of different classes moving along opposite directions such that the distances between classes can be enlarged. The promising results of MMSAE and LGCFL show that modification on the label space is effective. Fig. 2(c) and (d) show the corresponding precision-recall curves. We observe that compared with its several counterparts, our method still performs best.

Table 3 shows the MAP@R scores achieved by the compared methods and our method on NUS-WIDE dataset. From the table, we can see that the proposed method outperforms its counterparts significantly. The promising results of MMSAE can be attributed to its capability to preserve both original feature and semantic information. We also see that GMLDA, 3view-CCA, ml-CCA, JFSSL and our method perform better than CCA and PLS because of taking the semantic information into account. NUS-WIDE dataset is much larger than WIKI and MIRFLICKR datasets, so semantic information in NUS-WIDE provides much more relations to build the correspondence between data from different modalities. Fig. 2(e) and (f) show the corresponding precision-recall curves. MMSAE shows the best cross-modal retrieval performance at all recall levels.

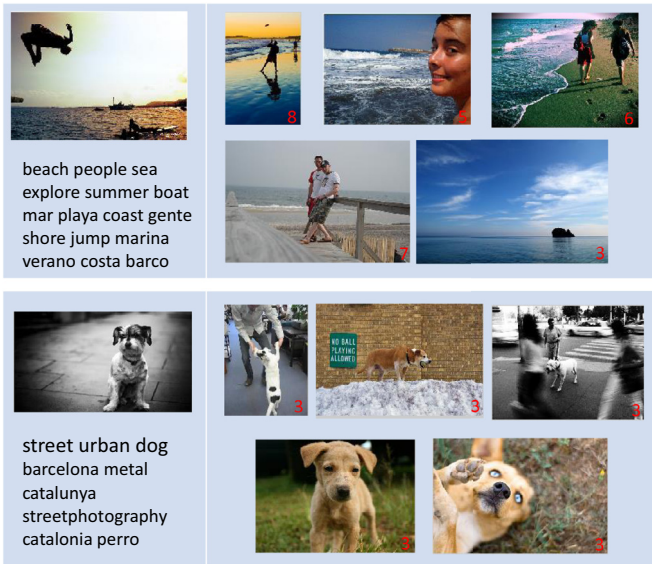
#### 4.6. Retrieved examples

To intuitively show the ranking performance of our method, we illustrate examples of the retrieved texts using image queries by MMSAE in Fig. 3 and examples of the retrieved images using text queries by MMSAE in Fig. 4. From the examples in Figs. 3 and 4, we can observe that our method can ensure that the top ranked examples are semantically consistent to the queries.

#### 4.7. Dimension of common space analysis

In LGCFL and JFSSL, the dimension of common space is fixed to be the same as the dimension of label space, while by using feature-aware label space dimension reduction paradigm we can choose the dimension of common space more flexibly for our method. In this subsection, we analysis the effect of dimension  $d$  of common space for cross-modal retrieval task on MIRFLICKR dataset. We change  $d$  in the range of {4, 8, 16, 32, 38} while fixing other parameters. The experimental results are shown in Fig. 5.

The experimental results show that our method performs well in a wide range of common space dimension. This indicates that the label space can be reduced without loss of information. As



**Fig. 4.** Two examples of text queries and the top 5 images retrieved by the proposed methods on MIRFLICKR dataset. First column contains the text queries and their corresponded images. Second column contains retrieved images. Red number indicates the number of overlapped labels to the query of each retrieved image.

lower dimension results in faster retrieval, it is useful to reduce the dimension of common space.

#### 4.8. Ablation study

We also verify the impact of different modules on the performance of MMSAE. Two variants are designed as baselines of

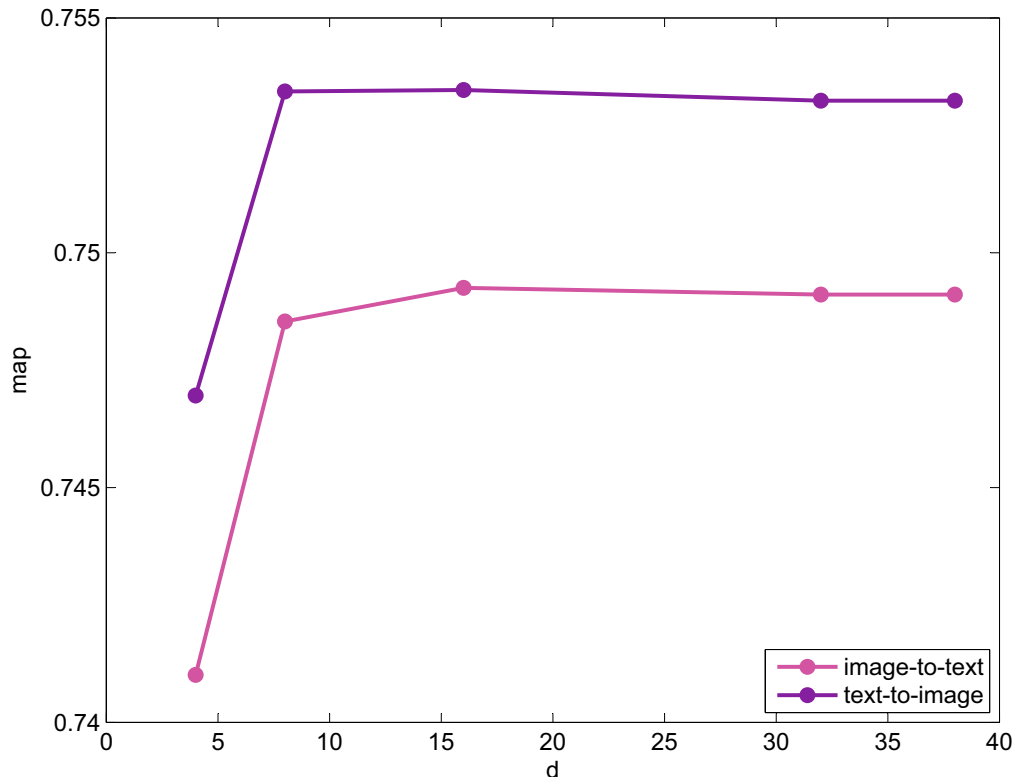
our model: (a) MMSAE-1 is built by removing the reconstruction terms, i.e., setting  $\alpha = 0$ ; (b) MMSAE-2 is built by removing the semantic preserving term, i.e., setting  $\beta = 0$ . Table 4 shows the comparison results on the WIKI and MIRFLICKR datasets. We can see from the table that our method can achieve a more accurate performance.

#### 4.9. Parameter sensitivity analysis

In this subsection, we analysis the effect of trade-off parameters  $\alpha$  and  $\beta$ .  $\alpha$  controls the reconstruction of original features and  $\beta$  controls keeping to the semantic code vector. We tune  $\alpha$  and  $\beta$  in the range of {10, 1, 0.1, 0.01, 0.001, 0.0001}. The experimental results on WIKI and MIRFLICKR datasets for image-to-text task and text-to-image task are shown in Fig. 6. We observe that the performance of our algorithm varies when the parameters are changing. The performance is more sensitive to  $\alpha$  than to  $\beta$ . Generally speaking, the proposed method obtains better performance when  $\alpha$  is in the range of 0.001–1.

#### 4.10. Convergence analysis

The iterative algorithm in Algorithm 1 is proposed to optimize the multi-modal semantic autoencoder objective. In Fig. 7, we plot the convergence curves of our iterative algorithm with respect to the loss value in Eq. (8) at each iteration on WIKI and MIRFLICKR datasets. We can see from the figures that the losses monotonically decrease at each iteration. With very few iterations, the losses become small and stable. Particularly, in our experiment, we ran the algorithm five iterations on all datasets.



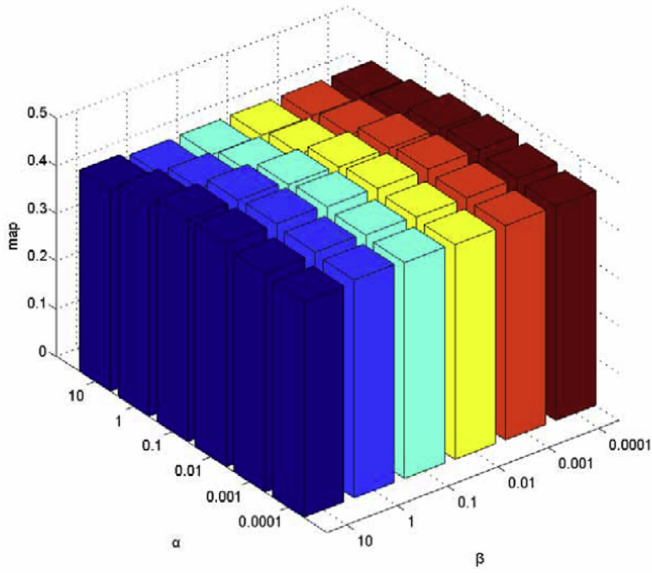
**Fig. 5.** Performance variation with respect to hidden dimension  $d$  on MIRFLICKR dataset.



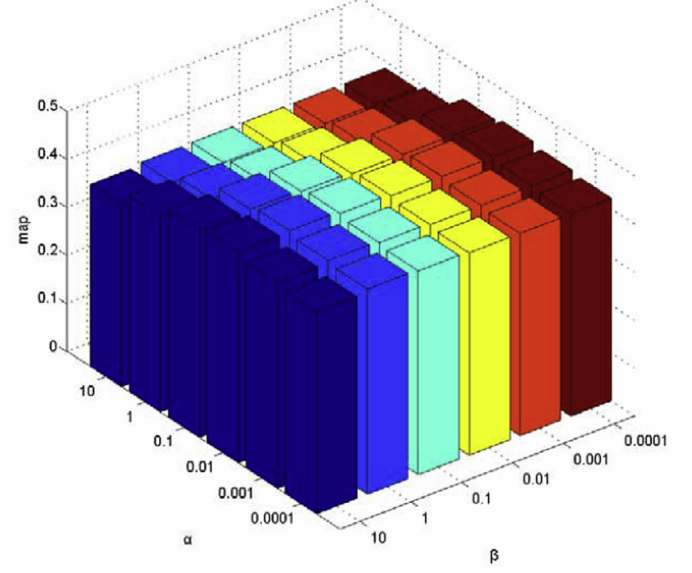
**Table 4**

The performance comparison with variants in terms of MAP@R.

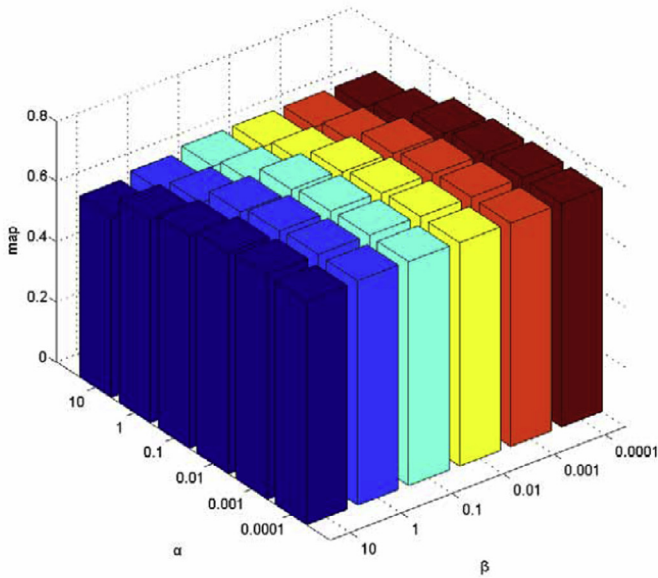
	Method	Task					
		$R = 50$			$R = all$		
		image-to-text	text-to-image	average	image-to-text	text-to-image	average
WIKI	MMSAE-1	0.472	0.573	0.523	0.457	0.426	0.442
	MMSAE-2	0.484	0.567	0.526	0.462	0.433	0.448
	MMSAE	<b>0.488</b>	<b>0.581</b>	<b>0.535</b>	<b>0.469</b>	<b>0.442</b>	<b>0.455</b>
MIRFLICKR	MMSAE-1	0.914	0.865	0.889	0.732	0.739	0.736
	MMSAE-2	0.908	0.862	0.885	0.739	0.741	0.740
	MMSAE	<b>0.920</b>	<b>0.884</b>	<b>0.902</b>	<b>0.749</b>	<b>0.753</b>	<b>0.751</b>



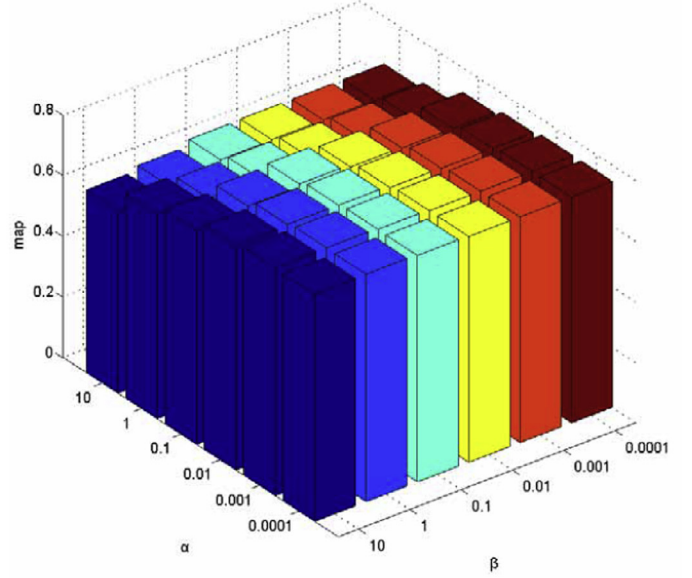
(a) image-to-text on WIKI



(b) text-to-image on WIKI



(c) image-to-text on MIRFLICKR



(d) text-to-image on MIRFLICKR

**Fig. 6.** Performance variation in terms of MAP@all with respect to  $\alpha$  and  $\beta$ .

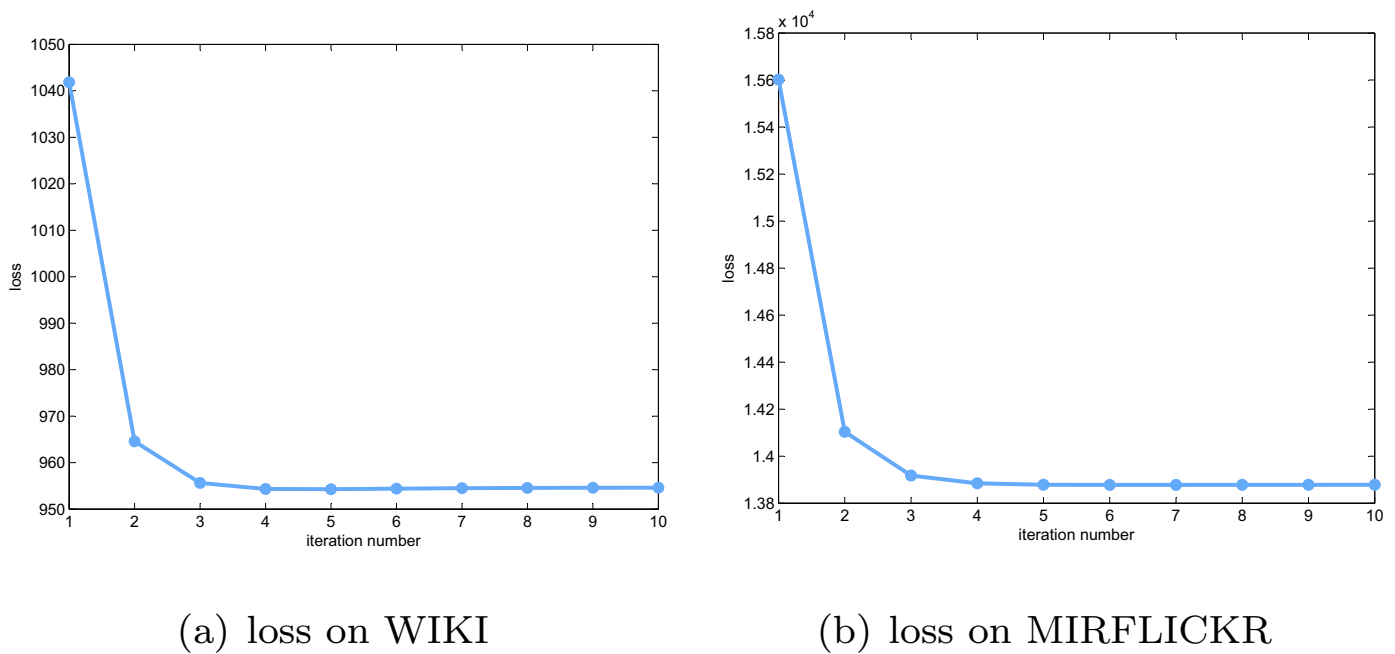


Fig. 7. Convergence curves of the objective function value in Eq. (8) using Algorithm 1.

## 5. Conclusion

In this paper, a novel method for the cross-modal retrieval task is proposed. In the proposed method, we learn modality-specific mappings to project data from different modalities to embeddings that preserve semantic information and original feature information in both modalities. We first learn feature-aware semantic code vectors which combine information from both feature spaces and label space. Then, we use encoder-decoder paradigm to learn projections which project image and text to the semantic code vector and simultaneously recover original features from the semantic code vector. The encoders and decoders are very simple linear functions which can be calculated efficiently. Experiments on three public datasets show that the proposed algorithm achieves the best performance compared to existing state-of-the-art algorithms.

## Acknowledgement

This work was supported in part by National Natural Science Foundation of China: 61672497, 61332016, 61620106009, 61650202 and U1636214, in part by National Basic Research Program of China (973 Program): 2015CB351802 and in part by Key Research Program of Frontier Sciences of CAS: QYZDJ-SSW-SYS013.

## References

- [1] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast uyghur text detector for complex background images, *IEEE Trans. Multimed.* 14 (8) (2017).
- [2] L. Zhuo, B. Cheng, J. Zhang, A comparative study of dimensionality reduction methods for large-scale image retrieval, *Neurocomputing* 141 (2014) 202–210.
- [3] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, *Neural Comput.* 16 (12) (2004) 2639–2664.
- [4] S. Abhishek, K. Abhishek, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2160–2167.
- [5] J. Masci, M.M. Bronstein, A.M. Bronstein, J. Schmidhuber, Multimodal similarity-preserving hashing, *IEEE Trans. Pattern Analysis and Machine Intelligence* 36 (4) (2014) 824–830.
- [6] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, PI-ranking: a novel ranking method for cross-modal retrieval, in: *Proceedings of the ACM on Multimedia Conference, ACM*, 2016, pp. 1355–1364.

- [7] Y. Wu, S. Wang, Q. Huang, Online asymmetric similarity learning for cross-modal retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4269–4278.
- [8] M. Xu, Z. Zhu, Y. Zhao, F. Sun, Subspace learning by kernel dependence maximization for cross-modal retrieval, *Neurocomputing* 309 (2) (2018) 94–105.
- [9] 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 ACM Multimedia*, 2010, pp. 251–260.
- [10] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2013, pp. 2088–2095.
- [11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.
- [12] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014, pp. 7–16.
- [13] E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3174–3183.
- [14] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2004, pp. 321–328.
- [15] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, *J. Mach. Learn. Res.* 7 (2006) 2399–2434. Nov.
- [16] H. Hotelling, Relations between two sets of variates, *Biometrika* 28 (3/4) (1936) 321–377.
- [17] 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) (2014) 210–233.
- [18] V. Ranjan, N. Rasiwasia, C. Jawahar, Multi-label cross-modal retrieval, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.
- [19] Y. He, S. Xiang, C. Kang, J. Wang, C. Pan, Cross-modal retrieval via deep and bidirectional representation learning, *IEEE Trans. Multimed.* 18 (7) (2016) 1363–1377.
- [20] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, *IEEE Trans. Multimed.* 17 (3) (2015) 370–381.
- [21] Y.-N. Chen, H.-T. Lin, Feature-aware label space dimension reduction for multi-label classification, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1529–1537.
- [22] C. Yan, L. Li, Cross-modality bridging and knowledge transferring for image understanding, *IEEE Trans. Multimed.* (99) (2018).
- [23] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, Y. Zhuang, Cross-media semantic representation via bi-directional learning to rank, in: *Proceedings of the ACM Multimedia*, 2013, pp. 877–886.
- [24] 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.

- [25] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Cross-modal retrieval using multi-ordered discriminative structured subspace learning, *IEEE Trans. Multimed.* 19 (6) (2017) 1220–1233.
- [26] Y. Zhen, D.Y. Yeung, A probabilistic model for multimodal hash function learning, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 940–948.
- [27] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2013, pp. 785–796.
- [28] X. Xu, L. He, A. Shimada, R.-i. Taniguchi, H. Lu, Learning unified binary codes for cross-modal retrieval via latent semantic hashing, *Neurocomputing* 213 (2016) 191–203.
- [29] D. Mandal, K.N. Chaudhury, S. Biswas, Generalized semantic preserving hashing for n-label cross-modal retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2633–2641.
- [30] D. Zhai, X. Liu, X. Ji, D. Zhao, S. Satoh, W. Gao, Supervised distributed hashing for large-scale multimedia retrieval, *IEEE Trans. Multimed.* 20 (3) (2018a) 675–686.
- [31] D. Zhai, X. Liu, H. Chang, Y. Zhen, X. Chen, M. Guo, W. Gao, Parametric local multiview hamming distance metric learning, *Pattern Recognit.* 75 (2018b) 250–262. Distance Metric Learning for Pattern Recognition.
- [32] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep Captioning with Multimodal Recurrent Neural Networks (m-rnn), *arXiv:1412.6632* (2014).
- [33] F. Feng, R. Li, X. Wang, Deep correspondence restricted boltzmann machine for cross-modal retrieval, *Neurocomputing* 154 (2015) 50–60.
- [34] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: *Proceedings of the ACM on Multimedia Conference*, ACM, 2017, pp. 154–162.
- [35] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [36] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [37] Y. Kim, Convolutional Neural Networks for Sentence Classification, *arXiv:1408.5882* (2014).
- [38] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: *INTERSPEECH*, (2014) 338–342.
- [39] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinomial data, in: *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2011, pp. 2407–2414.
- [40] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, K. Weinberger, Learning to rank with (a lot of) word features, *Inf. Retr.* 13 (3) (2010) 291–314.
- [41] D. Grangier, S. Bengio, A discriminative kernel-based approach to rank images from text queries, *Trans. pattern analysis and machine intelligence* 30 (8) (2008) 1371–1384.
- [42] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: *Proceedings of the 25th International Conference on Machine Learning*, ACM, 2008, pp. 1096–1103.
- [43] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, *Nature* 323 (6088) (1986) 533.
- [44] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, A graph regularized deep neural network for unsupervised image representation learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1203–1211.
- [45] Y.-l. Boureau, Y.L. Cun, et al., Sparse feature learning for deep belief networks, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2008, pp. 1185–1192.
- [46] R.H. Bartels, G.W. Stewart, Solution of the matrix equation  $ax + xb = c$  [4], *Commun. ACM* 15 (9) (1972) 820–826.
- [47] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022.
- [48] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: *Proceedings of the AAAI*, 2017, pp. 4278–4284.
- [49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: large-scale machine learning on heterogeneous distributed systems, *arXiv:1603.04467* (2016).
- [50] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ACM, 2008, pp. 39–43.
- [51] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: *Proceedings of the NIPS*, 2012, pp. 2222–2230.
- [52] 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 CIVR*, ACM, 2009, p. 48.
- [53] Z. Lin, G. Ding, M. Hu, J. Wang, Semantics-preserving hashing for cross-view retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3864–3872.
- [54] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, in: *Subspace, Latent Structure and Feature Selection*, Springer, 2006, pp. 34–51.



**Yiling Wu** received the B.S. degree in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2013. She is currently pursuing the Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her research interests include image and text retrieval, and deep learning.



**Shuhui Wang** received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in China, 2012. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include large-scale Web data mining, visual semantic analysis, and machine learning.



**Qingming Huang** received the B.S. degree in computer science and Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor and the Deputy Dean of the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China. He has authored over 300 academic papers in international journals, such as the *IEEE TRANSACTIONS ON IMAGE PROCESSING*, the *IEEE TRANSACTIONS ON MULTIMEDIA*, the *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, and top level international conferences, including the ACM Multimedia, International Conference on Computer Vision, Computer

Vision and Pattern Recognition, European Conference on Computer Vision, International Conference on Very Large Data Bases, and International Joint Conference on Artificial Intelligence. His current research interests include multimedia computing, image/video processing, pattern recognition, and computer vision.