Neurocomputing 355 (2019) 93–104

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Adversarial cross-modal retrieval based on dictionary learning

Fei Shang[a], Huaxiang Zhang[a,b,*], Lei Zhu[a,b], Jiande Sun[a,b]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan, 250014 Shandong, China
[b] Institute of Data Science and Technology, Shandong Normal University, Jinan, 250014 Shandong, China

ARTICLE INFO

Article history:
Received 2 December 2018
Revised 21 March 2019
Accepted 22 April 2019
Available online 10 May 2019

Communicated by Yongdong Zhang

Keywords:
Cross-modal retrieval
Adversarial learning
Dictionary learning

ABSTRACT

Existing cross-modal approaches focus on learning a subspace or using classical neural networks for similarity measurement of different modalities, which ignore the complex statistical properties of multimodal data. To settle the above problems, we propose a novel framework termed Adversarial Cross-Modal Retrieval Based on Dictionary Learning Algorithm (DLA-CMR). The dictionary learning serves as feature reconstructor to reconstruct discriminative features, while adversarial learning mines the statistical characteristics for each modality. Firstly, using all of the training (testing) samples to reconstruct each training (testing) sample, the specificity of each sample is maintained to some extent. Secondly, the weight of important features increases while that of secondary features decreases. This also makes the dimension of transformed visual modality approximate to textual modality. In addition, the adversarial learning guarantees that the transformed features maintain the inherent statistical characteristics of original features for each modality, and it requires transformed features to be statistically indistinguishable in common space. The transformed features must be maximally correlated to eliminate the heterogeneous gap. Comprehensive experimental results compared with 7 state-of-the-art methods on 4 widely-used datasets verify the effectiveness of our DLA-CMR method.

1. Introduction

In recent years, we have witnessed the vigorous development of multimodal data including videos, texts, 3D models, audio and images on the Internet, and also have perceived the transformation from single modality [1–4] to multimodal when describing things. As shown in Fig. 1, when we need to buy a filco mechanical keyboard on Amazon by. entering “filco” keyword in the search box, we are eager to retrieve all kinds of images describing the keyboard, display the 3D model of the keyboard from different sides, show the video of the keyboard’s usage and function, and even find the audio striking the keyboard from user evaluation. This advanced retrieval technique, which can present retrieval results with different modalities by a query of any modality, is called cross-modal retrieval. It helps us understand omnibearing information derived from different forms of data, such as texts, images, videos, 3D models, etc. The development of cross-modal retrieval [5–7] is imperative. Cross-modal retrieval can realize the fusion and supplement of various forms of information by analyzing the characteristic distribution of multimodal data in different modalities and utilizing the intrinsic semantic relations [8] between data, so as to achieve the goal of flexibly striding across various modalities according to semantic information.

While many researchers are working on effective representations of multimodal data and striving to narrow the heterogeneous gap between different modalities, the research has not been done well. With the rich semantic information of multimodal data, removing redundant information and utilizing strongly relevant and discriminative information attract more and more attention in cross-modal retrieval. Traditional dictionary learning approaches deal with the projection matrices using F norm and trace norm to realize features selection. In addition, graph related methods filter features employing $l_{2,1}$ norm and nuclear norm constraint. The appealing feature extraction methods are proposed to further explore features by multiple convolution layer. Although these methods learn discriminative information in a certain extent, they all ignore the correlation between samples, which can effectively maintain the inter-modal similarity relation.

Generally speaking, the text data is discrete, while the image representation is continuous. Considering the complicated statistical features of multimodal data, it is challenging to directly establish relationship between different modal data. The current mainstream approaches learn a common space for different modal data, which not only promote the similarity measure of multimodal data, but also ensure the largest relevance to transformed features.

[*] Corresponding author at: School of Information Science and Engineering, Shandong Normal University, Jinan, 250014 Shandong, China.
E-mail address: huaxzhang@hotmail.com (H. Zhang).

https://doi.org/10.1016/j.neucom.2019.04.041
0925-2312/© 2019 Elsevier B.V. All rights reserved.

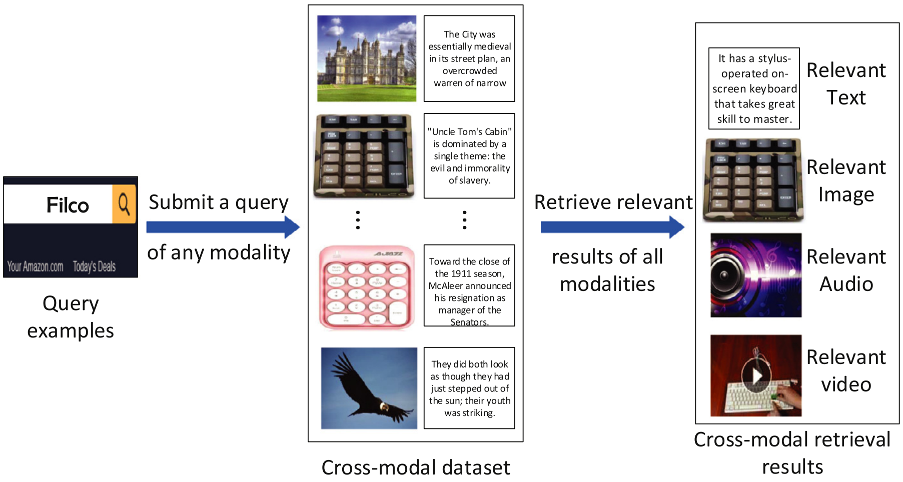

Filco
Your Amazon.com Today's Deals
Query examples
Submit a query of any modality
The City was essentially medieval in its street plan, an overcrowded warren of narrow
"Uncle Tom's Cabin" is dominated by a single theme: the evil and immorality of slavery.
Toward the close of the 1911 season, McAleer announced his resignation as manager of the Senators.
They did both look as though they had just stepped out of the sun; their youth was striking.
Cross-modal dataset
Retrieve relevant results of all modalities
It has a stylus-operated on-screen keyboard that takes great skill to master.
Relevant Text
Relevant Image
Relevant Audio
Relevant video
Cross-modal retrieval results

Fig. 1. An example of cross-modal retrieval with text, image, video and audio, which the input data and the retrieved data belong to different modalities.

However, these works show weaknesses that can not ensure the statistically indistinguishability of features following transformation.

In order to make use of strongly relevant and discriminative information and effectively deal with the complicated characteristics of multimodal data, the dictionary learning serves as feature reconstructor to reconstruct discriminant features, while adversarial learning mines the statistical characteristics for each modality. We propose a novel framework termed Adversarial Cross-Modal Retrieval Based on Dictionary Learning Algorithm (DLA-CMR). It aims to learn a common space for the image modality and text modality. The main contributions can be summarized as follows:

- **Discriminative features reconstruction by dictionary learning.** The original features extracted by CNN or SIFT are redundant, which may have a negative impact on cross-modal retrieval. The dimension of the image modality is much larger than that of the text modality. Projecting directly into a common space will cause a lot of information loss. So we use dictionary learning to reconstruct strongly relevant and discriminative information, which not only maintains the specificity of each sample to some extent, but also makes the dimension of transformed visual modality approximate to textual modality.
- **Cross-modal adversarial mechanism is established to maintain statistical characteristics of different modalities.** It consists of two competing processes acting as players in the minimax game, termed feature preserving and modality classification. Feature preserving keeps the statistical characteristics of the original characteristics of each modality in common space by label prediction and triplet constraints, which ensures the inter-modal invariance. Modality classification is equivalent to the discriminator in GAN, predicting the source modality of the transformed features. Feature preserving and modality classification are mutually antagonistic and reinforcing to learn a common space, under which the transformed features are statistically indistinguishable and the correlations between modalities are maximized.

The rest of this paper is organized as follows: We briefly review the existing cross-modal methods from two perspectives, including shallow learning methods and deep learning methods in Section 2. Our proposed DLA-CMR method is presented detailedly in Section 3. We show a variety of experimental results of DLA-CMR in the field of cross-modal retrieval in Section 4, and the conclusion of this paper will be given in Section 5.

2. Related work

Nowadays, the multimodal data in the network information have rapidly developed and have set the prairie ablaze. Exploration of cross-modal retrieval technology like a raging fire is carrying on. This method has been improved by experts from various fields, such as probability and statistics, graph theory and pattern recognition, etc. Cross-modal retrieval methods can be segmented into shallow learning methods [9,10] and deep learning methods [11,12]. The former brings forth the new through the old and plays an indispensable role in its application. Among shallow learning methods, graph related methods, dictionary learning methods have high accuracy and enjoy great popularity.

Graph related methods consist of graph based methods and graph regularization methods, which can describe intra- and interpair similarity and label relevance. By adopting various regularization on the projection matrices, these method get discriminative and robust information to some degree. Zhai et al. [13] proposed a novel algorithm called joint representation learning (JRL), which can fully excavate the pairwise correlation information as the supplement to the information of the original label. It utilizes $l_{2,1}$ norm to select sparse features, which is robust to the noise. Peng et al. [14] integrated different modality instances and their patches into a hypergraph. These patches can exploit local information and raise the variousness of training data. And then the high-order correlations between different modalities can be availably utilized from the hypergraph. Although these methods select better features, they all ignore the correlation between samples.

Dictionary learning methods are extraordinary popular in cross-modal retrieval, which utilize F norm and trace norm to learn sparse representation or collaborative representation for different modalities. Coupled dictionary learning in [15] extends the single type of media to cross-modal scene. By sparse coefficient mappings, the query data can be mapped into other modalities space. Bahrampour et al. [16] proposed an advance dictionary learning algorithm based task-driven. Multimodal dictionaries and classifiers corresponding to them are learned at the same time. Different from the traditional dictionary learning methods of using sparse representation, Shang et al. [17] proposed a novelty and effectiveness method which utilizes consistency preserving of collaborative representation. They carried out cross-modal retrieval in an isomorphic subspace which is projected by collaborative representation coefficients. The correlation over samples has been considered in the above methods, however there are some limitations in processing the excessive dimension differential in various modal characteristics.

With the prevalence of deep learning, the deep neural network has shown great vitality in the field of cross-media retrieval. It can excavate the potential complexity connections of cross-modal data in depth. Andrew et al. [18] presented a parameter model called

Deep Canonical Correlation Analysis (DCCA) which does not require reference training data to calculate the characteristic description of the testing samples in the shared space and has good scalability in computing complexity. DCCA contains two deep neural networks. By learning the multilevel nonlinear exchange, the output layer's correlation is the largest, so the isomorphic feature description of heterogeneous modality data is established.

Among numerous network models, Generative Adversarial Network (GAN) [19,20] is favored by many scholars for the reason that it does not require Markov chains and can generate high quality samples. It can be competent for real data generation tasks, such as image generation [21] and natural sentence generation. Reed et al. proposed an original deep architecture [22], which can automatically synthesize real images from the supplied text. It means converting from the visual concept of character level to picture element into reality.

Many researchers have contributed to the improvement of the network structure, however, the complicated statistical characteristics of multimodal data have not been well considered. Inspired by the reason that adding some explicit external information into original GAN can reduce training instability, Mirza and Osindero proposed conditional GANs (cGANs) [23] which provide additional information to guide network training. Radford et al. proposed a more steady GAN architecture called Deep Convolutional GAN (DCGAN) [24], which made use of batch normalization and different activation functions.

Of course, what interests us most is the application of GANs in cross-modal retrieval. MIPL laboratory take advantage of the current most popular generative adversarial network to imitate the joint distribution of different modalities. The generator and discriminator can fully excavate the information of intra-modality and inter-modality. GAN is also widely applied into semi-supervised and unsupervised learning for the reason that generative model can fill up missing information. Zhang et al. [25] structured a correlation graph which can seize the potential structures between multimodal data. They also utilized the triplet ranking loss to enhance the discriminant of the model. In their another article [26], they ingeniously used GAN to generate image-text pairs for unlabeled data. It expands datasets to improve retrieval accuracy. Peng et al. proposed an advanced and novel dual channel GAN model [27], which can take advantage of weight-sharing constraint to better capture multimodal semantic consistency and use two discriminators to identify simultaneously. These methods broadly enable the similarity measurement and semantic mining of multimodal data, while features obtained remains statistically separated. Wang et al. made feature projector and modality classifier serve as antagonist. This novel model called Adversarial Cross-Modal Retrieval (ACMR) [28] can effectively eliminate the heterogeneous gap. This method takes into account the statistical characteristics of multimodal data, nevertheless, the availability of discriminative features can not be guaranteed.

Inspired by the above works, we utilize dictionary learning to reconstruct discriminative features while using adversarial learning to mine the statistical characteristics for image and text. In our previous CR-CMR approach [17], the dimension of the reconstruction coefficient is $n_{tr} \times n_{tr}$, and is irrelevant to the original features. It is only associated with the instance number of the training set. However, in our DLA-CMR approach, $V \in R^{n_{tr} \times 4096}$ and $T \in R^{n_{tr} \times 5000}$ respectively represent the transposition of coefficients A_V and A_T. Both of them are not only take into account the particularity of each instance, but also make the transformed visual modal dimension approximate to textual modality. Comparing with ACMR, we fully consider the effect of the number of full-connected layers and nodes in each layer on different modalities when we construct a common space. We deploy four-layer full-connected neural networks for image modality and three-layer full-connected neural networks for text modality. In addition, we do co-optimization of dictionary learning and adversarial learning to learn optimal parameters.

3. The proposed approach

The overall architecture of our proposed DLA-CMR approach is illustrated in Fig. 2. DLA-CMR consists of two parts, called dictionary learning model [29,30] and adversarial learning model [31,32]. The former utilizes the reconstruction coefficients instead of the original features, which plays the role of feature reconstruction. It not only improves the accuracy of retrieval, but also accelerates the speed of retrieval. The latter consists of feature preserving and modality classification. Among feature preserving, the triplet ranking is taken into account to reduce inter-modal invariance loss, simultaneously maintaining the minimum distance between the same class and the maximum distance between different class. In addition, the feature discrimination can reduce intra-modal discrimination loss, by which the semantic labels of transformed features can be predicted. Different from traditional GANs which distinguish whether data is generated or real, our modality classification aims to distinguish the original modality of data. In order to ensure that the transformed features are statistically indistinguishable and maximally correlated in common space, feature preserving and modality classification are mutually antagonistic and reinforcing.

3.1. Notation

The whole DLA-CMR approach includes dictionary learning model and adversarial learning model. We first introduce the formal definition in detail. We carry out the experiments on four widely-used cross-modal datasets, which consist of image modality and text modality, namely X and Y, respectively. Let $D = \{D_{tr}, D_{te}\}$ represents the cross-modal dataset, where D_{tr} represents the training set and D_{te} represents the testing set. In particular, $D_{tr} = \{X_{tr}, Y_{tr}\}$ consists of m image-text pairs with each data pair $d_i = (x_i, y_i)$ including a d_v dimensional visual feature vector x_i and d_t dimensional textual feature vector y_i. Let $X_{tr} = \{x_1, \ldots, x_{n_{tr}}\} \subset R^{d_v \times n_{tr}}$, $Y_{tr} = \{y_1, \ldots, y_{n_{tr}}\} \subset R^{d_t \times n_{tr}}$ represent visual feature matrix, textual feature matrix, respectively. Similarly, $D_{te} = \{X_{te}, Y_{te}\}$, where $X_{te} = \{x_1, \ldots, x_{n_{te}}\} \subset R^{d_v \times n_{te}}$, $Y_{te} = \{y_1, \ldots, y_{n_{te}}\} \subset R^{d_t \times n_{te}}$. n_{tr} and n_{te} represent the number of training set and testing set, respectively.

3.2. Dictionary learning model

Dictionary learning model can learn appropriate dictionaries for original dense images and texts, and encode them into corresponding reconstruction representations, thus effectively improving retrieval accuracy. The contributions of dictionary learning model are as follows: Firstly, using all of the training (testing) samples to reconstruct each training (testing) sample, the specificity of each sample is maintained to some extent. Second, the reconstruction coefficients play the role of feature reconstruction instead of the original features. The weight of important features increases while that of secondary features decreases. Finally, dictionary learning also makes the dimension of transformed visual modality approximate to the textual modality.

We assume reconstruction coefficients $A_V \subset R^{k_1 \times n_{tr}}$ and $A_T \subset R^{k_2 \times n_{tr}}$ take the place of original features X_{tr} and Y_{tr}, respectively. Let $D_V \subset R^{d_v \times k_1}$ and $D_T \subset R^{d_t \times k_2}$ represent dictionaries corresponding to image and text. k_1 and k_2 are the size of visual dictionary and textual dictionary, respectively, where we set 4096 and 5000 in our experiments. $f(\cdot)$ is defined as the correlation between image

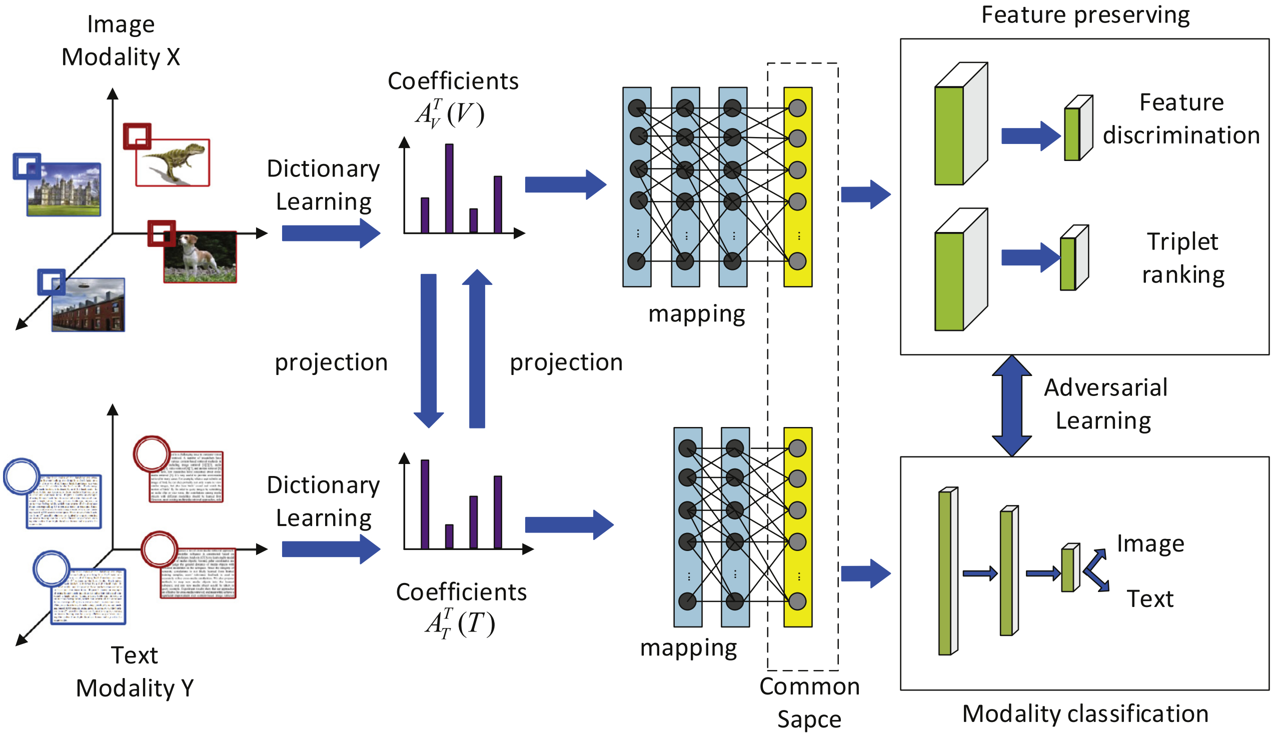


Fig. 2. The overall architecture of our proposed DLA-CMR approach.

and text. The dictionary learning model can formally be formulated as:

$$\begin{aligned} &\mathop{\arg\min}_{D_V, D_T, A_V, A_T} \|X - D_V A_V\|_F^2 + \|Y - D_T A_T\|_F^2 \\ &\quad + \lambda_1 (\|A_V\|_F^2 + \|A_T\|_F^2) + f(A_V, A_T) \\ &s.t. \sum_{i=1}^{d_V} D_{V_{i,j}}^2 \leq c, \sum_{i=1}^{d_T} D_{T_{i,k}}^2 \leq c, \\ &\forall j = 1, \ldots, 4096, \forall k = 1, \ldots, 5000. \end{aligned} \tag{1}$$

Then the definition of $f(A_V, A_T)$ can be formulated as:

$$\begin{aligned} f(A_V, A_T) = \lambda_2 (\|A_V - P_1 A_T\|_F^2 + \|A_T - P_2 A_V\|_F^2) \\ + \lambda_3 (\|P_1\|_F^2 + \|P_2\|_F^2), \end{aligned} \tag{2}$$

where P_1 and P_2 are the projection matrices, and the first term is the projection fidelity term which indicates the projection error between reconstruction coefficients A_V and A_T.

Above all, we learned dictionaries D_V and D_T from the training set.

$$D_V = (X A_V^T)(A_V A_V^T + \Lambda)^{-1} \tag{3}$$

$$D_T = (Y A_T^T)(A_T A_T^T + \Lambda)^{-1} \tag{4}$$

where Λ represents diagonal matrix. Please refer to our previous work [17] for specific solutions and the values of parameters λ_1, λ_2, λ_3.

Given testing set $D_{te} = \{X_{te}, Y_{te}\}$, we can use D_V and D_T to restructure A_V^{te} and A_T^{te}, respectively.

$$\mathop{\arg\min}_{A_V^{te}} \|X_{te} - D_V A_V^{te}\|_F^2 + \lambda_1 \|A_V^{te}\|_F^2, \tag{5}$$

$$\mathop{\arg\min}_{A_T^{te}} \|Y_{te} - D_T A_T^{te}\|_F^2 + \lambda_1 \|A_T^{te}\|_F^2. \tag{6}$$

We can get the reconstruction coefficients A_V^{te} and A_T^{te} as following:

$$A_V^{te} = (D_V^T D_V + \lambda_1 E)^{-1} (D_V^T X_{te}) \tag{7}$$

$$A_T^{te} = (D_T^T D_T + \lambda_1 E)^{-1} (D_T^T Y_{te}) \tag{8}$$

The transposition of coefficients A_V and A_T are renamed V and T. Similarly, the transposition of A_V^{te}, A_V^{te} are used as image testing set and text testing set respectively in the next model.

3.3. Adversarial learning model

Adversarial learning model is regarded as a minimax game involving two plays which have opposite training goals. Specially, these players are mutually antagonistic and reinforcing to reach the Nash Equilibrium.

3.3.1. Modality classifier

In order to guarantee transformed features to be statistically indistinguishable, we introduce a modality classifier with predicting the source of transformed features, which corresponds to the discriminator in traditional GAN. It achieves the purpose of promoting inter-modal invariance by minimizing the classification error rate. It marks the image (text) morphological feature as the label 01(10), which is implemented by a three-layer feed-forward neural network with parameter θ_A. The adversarial loss can formally be formulated as:

$$L_{adv}(\theta_A) = -\frac{1}{n} \sum_{i=1}^{n} (m_i \cdot (\log D(v_i; \theta_A) + \log(1 - D(v_t; \theta_A))), \tag{9}$$

Where we utilize cross-entropy to represent optimization functions, which avoids gradient disappearance to some extent. Furthermore, m_i is a one-hot vector which represents the ground-truth modality label of each instance, while n represents the number of instances.

3.3.2. Triplet ranking

Taking image retrieval text as an example, we consider the following two measures to maintain statistical similarity of transformed features: (1) image and corresponding positive text, termed

$\{(v_i, t_i^+)\}_i$ (2) image, corresponding positive text and negative text, termed $\{(v_i, t_j^+, t_k^-)\}_i$. The former is a general formula of pairwise learning [33,34], and the latter is the triplet ranking method adopted in this paper. While considering the minimum distance between the same class, the maximum distance between different class is taken into account. We use $f_V(V;\ \theta_V)$ and $f_T(T;\ \theta_T)$ represent the transformed features in common space, and all distances can be computed as follows:

$$l_2(v,t) = \|f_V(v;\theta_V) - f_T(t;\theta_T)||_2. \tag{10}$$

And then, the inter-modal invariance loss can use the following definition respectively:

$$L_{tri}V(\theta_V) = \sum_{i,j,k} \max(0, l_2(v_i, t_j^+) - l_2(v_i, t_k^-) + \mu), \tag{11}$$

$$L_{tri}T(\theta_T) = \sum_{i,j,k} \max(0, l_2(t_i, v_j^+) - l_2(t_i, v_k^-) + \mu), \tag{12}$$

where u is a balance parameter. Besides, the DNNs have abundant parameters w_v^l and w_t^l in each layer, we add the following regularization terms to avoid parametric overfitting.

$$L_{reg} = \sum_{l=1}^{l} (\|w_v^l||_F + \|w_t^l||_F). \tag{13}$$

Finally, the full feature preserving loss can be defined as the combination of the following:

$$L_{fr}(\theta_V, \theta_T, \theta_D) = \alpha(L_{tri}V(\theta_V) + L_{tri}T(\theta_T)) + \beta L_{dis}(\theta_D) + L_{reg}, \tag{14}$$

where α and β are hyper-parameters.

3.3.3. Feature discrimination

In order to guarantee that the transformed features are still discriminative within the modality, we propose a feature discriminator, which can predict semantic labels regardless of the source modality of the feature is text or image. It takes one transformed feature as input and outputs the probability distribution $\hat{p}_i$ of its semantic labels. Similar to the definition in modality classifier, we also use cross-entropy to formally define the loss function as follows:

$$L_{dis}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(l_i \cdot (\log \hat{p}_i(v_i) + \log \hat{p}_i(t_i))). \tag{15}$$

3.3.4. Optimization

On account of the opposition between the optimization goal of feature preserving and modality classification, we utilize minimax game combining the above three loss items to define the objective function. The overall objective function of adversarial learning can be formally defined as follows:

$$(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_D) = \underset{\theta_V, \theta_T, \theta_D}{\arg\min}(L_{fr}(\theta_V, \theta_T, \theta_D) - L_{adv}(\theta_A)), \tag{16}$$

$$\hat{\theta}_A = \underset{\theta_A}{\arg\max}(L_{fr}(\theta_V, \theta_T, \theta_D) - L_{adv}(\theta_A)). \tag{17}$$

It is optimized through stochastic gradient descent (SGD) algorithms which can effectively excavate information, and achieve significant effect in the early iteration. We utilize Gradient Reversal Layer (GRL) to guarantee min-max optimization can be implemented simultaneously. The co-optimization of dictionary learning and adversarial learning as shown in the Algorithm 1.

3.4. Implementation details

Our proposed DLA-CMR approach is implemented by **Tensorflow**, which is the most popular scientific computing framework. We will specifically introduce the implementation details of common space, feature discrimination and modality classifier in this paragraph.

(1) Common space: Four full-connected layers are adopted in image modality and three full-connected layers are adopted in the text modality when map original modalities from nonlinear high-level space into a common space, and a *tah* activation function layer follows each full-connected layer so that every value can be projected to $[-1, 1]$. The dimension of a common space is equal to the integral number of semantic categories. As we all know, there are 10, 20 and 100 categories on Wikipedia dataset, Pascal Sentence dataset and INRIA-Websearch dataset, respectively. So we choose 100-dimension as the dimension of the common space. In conclusion, we deploy $4096 \rightarrow 2000 \rightarrow 512 \rightarrow 200 \rightarrow 100$ for image modality and $5000 \rightarrow 512 \rightarrow 200 \rightarrow 100$ for text modality.

(2) Feature discrimination: The feature discrimination utilizes the transformed features of image-text pairs from the common space as input and a probability distribution of semantic class per instance as output. Different from the common space, we use *softmax* activation function in feature discrimination process.

(3) Modality classifier: This classifier uses an unknown transformed feature from common space as input and output probability of every modality. In order to identify the original modality of features, three full-connect layers and softmax activation function are used in modality classifier. And we deploy $f \rightarrow 50 \rightarrow 2$ for modality classifier.

We conduct several experiments to validate the robustness of our proposed DLA-CMR approach in this paragraph. The specific details are as follows:

The batch size varies along with datasets, we set to 64 on the first three datasets and 128 on the last dataset. α and β are weight coefficients which represent the feature discrimination loss and triplet ranking loss, respectively. In experiments, we tune α and β from {0.0001, 0.001, 0.01} and {10, 100, 1000}, respectively. As shown in the Fig. 3(a), grid search is performed for them by fixing λ at 0.01 on Wikipedia-CNN dataset. And the analysis of μ is shown in Fig. 3(b), we tune μ from {0.0001, 0.001, 0.01, 0.02, 0.1}. Finally the parameters α, β and μ are set to 100, 0.02, 0.0001 on Wikipedia dataset, 200, 0.002, 0.02 on Wikipedia-CNN dataset, 100, 0.02, 0.01 on Pascal Sentence dataset and 100, 0.02, 0.01 on INRIA-Websearch dataset, respectively.

4. Experiments

In this section, we exhibit the experimental results of our proposed DLA-CMR method for two cross-modal retrieval tasks. Firstly, we introduce four widely-used datasets, evaluation metrics, parameter tuning and implementation details. And then we compare the results of DLA-CMR with 5 state-of-the-art methods and 2 advanced DNNs methods on four datasets. Meanwhile, the further analysis of DLA-CMR is displayed in the end of this section.

4.1. Datasets

We briefly introduce four datasets used in our experiments, which are Wikipedia dataset, Wikipedia-CNN dataset, Pascal Sentence dataset and INRIA-Websearch dataset. The first three datasets are often seen in the filed of cross-modal literature. The last is a relatively large-scale image/text dataset which is constructed by Wei et al. [35]. Table 1 summarizes the general statistics of the four datasets.

```
Input: Image training set $X_{tr}$; Text training set $Y_{tr}$; Image testing set $X_{te}$; Text testing set $Y_{te}$; batch size $b_s$; parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\alpha$, $\beta$, $\mu$, $\lambda$
Initialize $D_V$, $D_T$ , $A_V$, $A_T$, $P_1$, $P_2$ as random matrices for iteration $i$=0.
Repeat until convergence:
    1:Fix $D_V$, $D_T$, $P_1$, $P_2$, update $A_V$, $A_T$.
    2:Fix $P_1$, $P_2$, $A_V$, $A_T$, update $D_V$, $D_T$.
    3:Fix $D_V$, $D_T$, $A_V$, $A_T$, update $P_1$, $P_2$ .
    4:Set $i$=$i$ + 1.
Obtain Coefficients $V = A_V^T$, $T = A_T^T$.
Utilize $D_V$, $D_T$, get $A_V^{te}$, $A_T^{te}$.
Repeat until convergence:
    1:for k steps do
        2:update parameters $\theta_V$, $\theta_T$, $\theta_D$ by descending their
        stochastic gradients:
        3: $\theta_V \leftarrow \theta_V - \mu \cdot \nabla_{\theta_V} \frac{1}{b_s}(L_{fr} - L_{adv})$
        4: $\theta_T \leftarrow \theta_T - \mu \cdot \nabla_{\theta_T} \frac{1}{b_s}(L_{fr} - L_{adv})$
        5: $\theta_D \leftarrow \theta_D - \mu \cdot \nabla_{\theta_D} \frac{1}{b_s}(L_{fr} - L_{adv})$
    6:end for
    7:update parameters $\theta_A$ by ascending its stochastic gradeints through GRL:
    8: $\theta_A \leftarrow \theta_A + \mu \cdot \lambda \cdot \nabla_{\theta_A} \frac{1}{b_s}(L_{fr} - L_{adv})$
Output: transformed features in common space.
```

Algorithm 1. Pseudocode of optimizing our DLA-CMR.

Table 1
General statistics of datasets used in our experiments.

Datasets	Wikipedia	Wikipedia-CNN	Pascal Sentence	INRIA-Websearch
Database	2866	2866	1000	14698
Query	693	693	400	4366
Training	2173	2173	600	10332
Visual Feature	SIFT (128-D)	CNN (4096-D)	CNN (4096-D)	CNN (4096-D)
Text Feature	LDA (10-D)	LDA (100-D)	LDA (100-D)	LDA (1000-D)

Wikipedia-CNN dataset is generated by Wikipedia's featured article, including 2173 image-text pairs for training and 693 image-text pairs for testing. It is divided into 10 different semantic categories and each pair of data is labeled as one of 10 semantic categories. Each image is denoted with 4096-dimension CNN visual features and each text is denoted with 100-dimension LDA textual features. By contrast, the number of **Wikipedia** dataset is same as Wikipedia-CNN dataset, while image and text are extracted with 128–dimensional SIFT feature and 10-dimensional LDA feature, respectively.

Pascal Sentence dataset contains 20 semantic categories of data, each semantic categories has 50 image-text pairs. 60 per cent of image-text pairs from each category is randomly selected for training, and the rest for testing. For image representation, 4096-

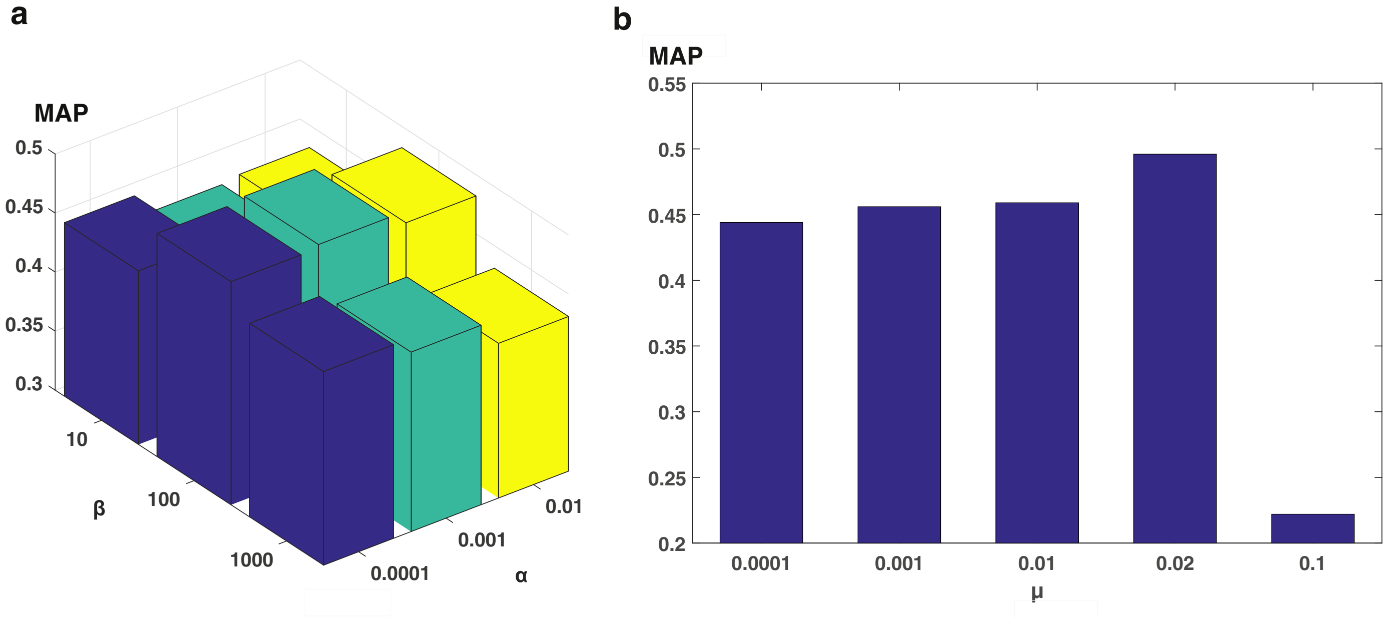


Fig. 3. Experimental results of DLA-CMR with different values of model parameters: (a)α and β; (b) μ on Wikipedia-CNN dataset.

dimension CNN visual features are used to describe one image. For text representation, 100-dimension LDA textual features are used to describe one text.

INRIA-Websearch is a relatively large-scale image/text dataset with 14698 instances belonging to 100 categories in our experiments. Each image-text pair corresponding to a semantic label. We take 70/30 per cent instances in each class as the training/testing set. 4096-dimensional CNN feature (1000-dimensional LDA textual feature) is used for image (text) representation.

4.2. Retrieval tasks and evaluation metrics

We perform bi-modal retrieval in our experiments which consists of following two sub-tasks:

Image query text (termed I2T): Given one image as query, to retrieve the corresponding text in the testing set by similarity matrix.

Text query image (termed T2I): Given one text as query, to retrieve corresponding image in the testing set by similarity matrix. We also calculate their average value in the experiments. In addition, the steps of cross-modal retrieval are as follows: (1) Training set is used to common representation learning which can learn a common space for the image modality and text modality. (2) Projecting the testing set into common space by the learned projection matrix. And (3) Sorting the similarity matrix to perform cross-modal retrieval.

In particular, we utilize three widely-used evaluation metrics to measure retrieval performance: Precision-Scope curve reveals the performance variations concerning the number of retrieved instances, Mean Average Precision (MAP) which are obtained by computing the mean of average precision (AP), and MAP for each class, concrete definition is as follows:

$$AP = \frac{1}{N}\sum_{r=1}^{R} prec(r)\delta(r) \tag{18}$$

where N is the number of the relevant instances in database, $prec(r)$ represents the accuracy of the r retrieved instances. If the k-th instance is relevant with query term $\delta(r)=1$ otherwise $\delta(r)=0$.

4.3. Compared methods

Three-view CCA (CCA-3V/T-V CCA) [36] is an original CCA variant, which incorporates a semantic class view to capture high-level image semantics. So the three-view refers to semantic class view, image view and text view. It utilizes explicit nonlinear kernel to improve the accuracy of cross-modal retrieval tasks.

Joint feature selection and subspace learning (JFSSL) [37] is a patulous version of LCFS [38], which studies the projection matrices for different modalities, and solves the problem of similarity measurement. By carrying out l_{21} norm constraint on projection matrices, more discriminative features are selected.

Modality-dependent cross-media retrieval (MDCR) [35] is a classical task-specific approach, which aims at learning two couples of projection matrices for I2T task and T2I task respectively rather than one same projection matrix for all tasks.

Joint latent subspace learning and regression (JLSLR) [39] combines spectral regression method and graph model to jointly learn the regress and latent space. The regression error for multimodal is minimized.

Generalized Semi-supervised Structured Subspace Learning (GSS-SL) [40] first uses label graph constraints to predict class labels for unlabeled data, and then uses different class labels that can reflect semantic information to bridge different modalities. It divides the training set into labeled data and unlabeled data in a certain proportion, and tries different proportions.

Cross-media Retrieval with collective deep semantic learning (CR-CDSL) [41] starts off the global structure information of multimodal data and can excavate latent semantic information in unlabeled instances. Weak semantic labels of unlabeled instances are marked as strong semantic labels to enhance the classification discrimination ability and semantic modeling ability of retrieval model, and a more meaningful semantic space can be found.

Adversarial cross-modal retrieval (ACMR) [28] is typical adversarial learning method which could learn modality-invariant and discriminative representations of different modalities. Modality classifier and feature projector compete against each other so that a couple of better feature representations are obtained.

4.4. Experiment results

The experiment results are evaluated in terms of Precision-scope curves, MAP score and MAP for each class are shown in Tables 2, 3 and Figs. 4, 5. Our compared methods consists of five shallow learning methods and two advanced deep learning meth-

Table 2
MAP of all compared approaches on Wikipedia dataset and Wikipedia-CNN dataset. The best result in each column is marked with bold. (D) represents deep learning methods.

Methods	Wikipedia			Wikipedia-CNN		
	I2T	T2I	Avg	I2T	T2I	Avg
CCA-3V	0.228	0.205	0.217	0.311	0.316	0.314
JFSSL	0.306	0.228	0.267	0.369	0.323	0.346
MDCR	0.287	0.225	0.256	0.422	0.382	0.402
JLSLR	0.236	0.212	0.224	0.394	0.369	0.382
GSS-SL	-	-	-	0.424	0.384	0.404
CR-CDSL(D)	0.348	0.249	0.299	0.508	0.443	0.475
ACMR(D)	0.351	0.245	0.298	0.506	0.428	0.467
DLA-CMR	**0.369**	**0.261**	**0.315**	**0.539**	**0.453**	**0.496**

Table 3
MAP of all compared approaches on Pascal Sentence dataset and INRIA-Websearch dataset. The best result in each column is marked with bold. (D) represents deep learning methods.

Methods	Pascal Sentence			INRIA-Websearch		
	I2T	T2I	Avg	I2T	T2I	Avg
CCA-3V	0.337	0.439	0.388	0.329	0.501	0.415
JFSSL	0.449	0.473	0.461	0.533	0.563	0.548
MDCR	0.447	0.451	0.449	0.521	0.551	0.535
JLSLR	0.454	0.455	0.455	0.525	0.545	0.535
GSS-SL	0.468	0.464	0.466	0.531	0.557	0.544
CR-CDSL(D)	-	-	-	0.559	0.608	0.584
ACMR(D)	0.468	0.501	0.485	0.558	0.649	0.604
DLA-CMR	**0.498**	**0.546**	**0.522**	**0.561**	**0.656**	**0.609**

ods. - represents that this compared method does not work on the corresponding dataset. The experimental results of all the compared methods are obtained from the source codes provided by their authors. From these experiment results, we can clear that our DLA-CMR achieves the best performance than seven compared methods on sub-retrieval tasks. The specific performance evaluation is as follows:

The MAP scores of two sub-tasks and the average MAP score on Wikipedia dataset and Wikipedia-CNN dataset are shown in Table 2. The average MAP score of I2T task and T2I task has been improved from 0.299 (0.467) to 0.315 (0.496) on Wikipedia (Wikipedia-CNN) dataset. Among seven compared methods, two advanced deep learning methods achieve better performance than five traditional shallow learning methods, where ACMR achieves the highest accuracy in these compared methods. It is indicated that neural networks have better performance than traditional machine learning models. Besides, we can see that the same method on Wikipedia-CNN significantly outperforms Wikipedia dataset. The improvement of performance due to CNN visual feature has stronger expression than SIFT feature.

Specific analysis of the MAP scores on Pascal Sentence dataset and INRIA-Websearch dataset is shown in Table 3. On the Pascal Sentence dataset, we can observe that our DLA-CMR obtains the best MAP score of 0.498 on I2T, 0.546 on T2I and 0.522 on their average value. Compared with the best shallow learning method GSS-SL, our DLA-CMR obtains an inspiring precision improvement from 0.468 to 0.498 on I2T, and from 0.464 to 0.546 on T2I task. Even compared with deep learning method ACMR, our approach increases 3%, 4.5% and 3.7%, respectively. The trend on INRIA-Websearch dataset is similar to Pascal Sentence dataset according to Table 3. We also give some successful analyses on our DLA-CMR as follows:

Firstly, CCA-3V, JFSSL, MDCR, JLSLR and GSS-SL are supervised methods which exploit semantic information to learn feature representation. In particular, CCA-3V utilizes explicit nonlinear kernel to improve the accuracy of cross-modal retrieval tasks. MDCR is a classical modal-independent method. JFSSL, JLSLR and GSS-SL use graph regularization constraints on the basis of semantic information. But our DLA-CMR is significantly superior to above state-of-the-art methods. The reason for this is that DLA-CMR utilizes dictionary learning to enhance the weight of important features, and uses adversarial learning to obtain more discriminative feature representation. Besides the dimension of transformed visual modality approximate to textual modality, which is conducive to learning common space.

Second, five shallow learning methods and CR-CDSL all take advantage of pairwise loss. They only exploit the information of image-text pairs, ignoring the information of different semantic category data. While considering the minimum distance between the same class, the maximum distance between different classes is not taken into account. Our DLA-CMR significantly outperforms the methods based pairwise loss. It indicates that triplet ranking has powerful ability to perform the feature correlation analysis.

Finally, ACMR does not use feature reconstruction before adversarial learning, which can not learn a better common space for image modality and text modality. Our DLA-CMR outperforms it on above four datasets. As the weight of important features increases, the weight of secondary features reduces after dictionary learning.

The MAP for each class on Wikipedia-CNN dataset and Pascal Sentence dataset is shown in Fig. 4. From Fig. 4(a), (c) and (e), we can see that all the methods achieve well performance on “biology”, “sport&recreation” and “warfare” class on Wikipedia-CNN dataset. The reason is that these three classes have more obvious characteristics than other classes. To the contrary, methods achieve the worst performance on “art&architecture” class due to its confusing features. Similarly, it can be observed from Fig. 4(b), (d) and (f) that some classes have high-level semantics, such as “plane”, “cat” and “train” class, which may lead to more discriminative characteristics when carrying out cross-modal retrieval. Our DLA-CMR approach achieves the best accuracy on most classes on both Wikipedia dataset and Pascal Sentence dataset, which validates the effectiveness of our approach.

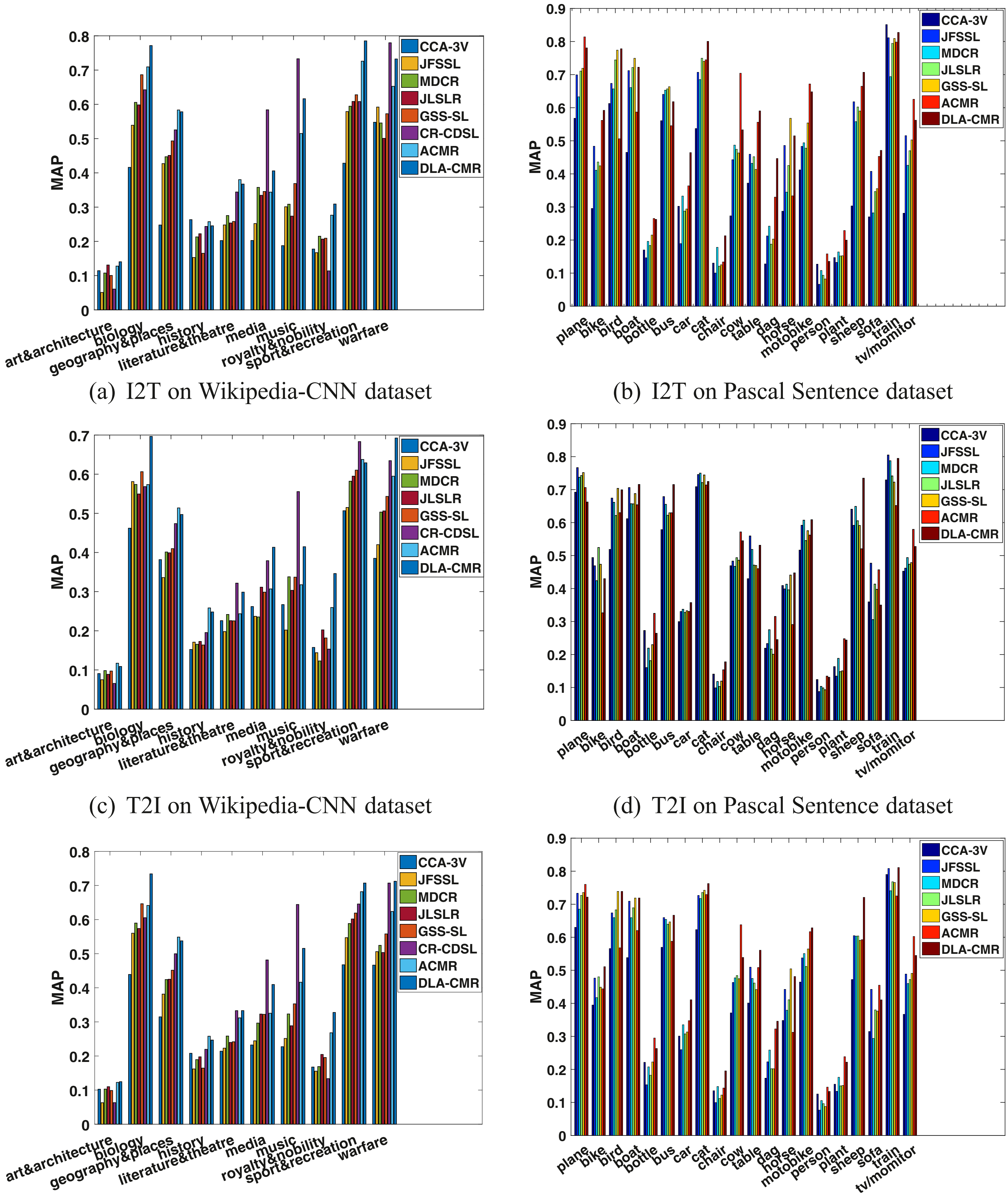
The Precision-Scope curves of T2I task and T2I task are shown in Fig. 5(a) and (b). We compare our DLA-CMR approach with the best shallow learning methods JFSSL and the best deep learning methods ACMR on Wikipedia dataset. We can get that the MAP scores correspond to the precision-scope curves with K ranging from 50 to 1000. And our DLA-CMR outperforms its compared methods.

4.5. Further analysis on DLA-CMR

In order to further verify the effectiveness of each component in our proposed DLA-CMR approach, five variants of DLA-CMR are conducted and compared with each other. The performance of our DLA-CMR and its four variants is shown in Tables 4 and 5. Specially, “DL” and “AL” refer to dictionary learning and adversarial learning, respectively. The detailed analysis is as follows:

(1) Performance of adversarial learning model: The adversarial learning model consists of feature preserving and modality classification as mentioned above, while feature preserving includes feature discrimination and triplet ranking. To demonstrate the contribution on adversarial learning model, we conduct four variants and calculate their MAP scores as shown in Table 4. “tr” denotes the triplet ranking and ”fd” denotes the feature discrimination.

We can observe that DLA-CMR is significantly superior to “DLA-CMR without AL”. Because dictionary learning can not maintain the statistical characteristics of the original feature for each modality in common space. “DLA-CMR without tr” uses dictionary learning, feature discrimination and modality classifier. It is inferior to DLA-CMR because triplet ranking can maintain inter-modal invariance. ”DLA-CMR without fd” uses dictionary learning, triplet ranking and



(a) I2T on Wikipedia-CNN dataset

(b) I2T on Pascal Sentence dataset

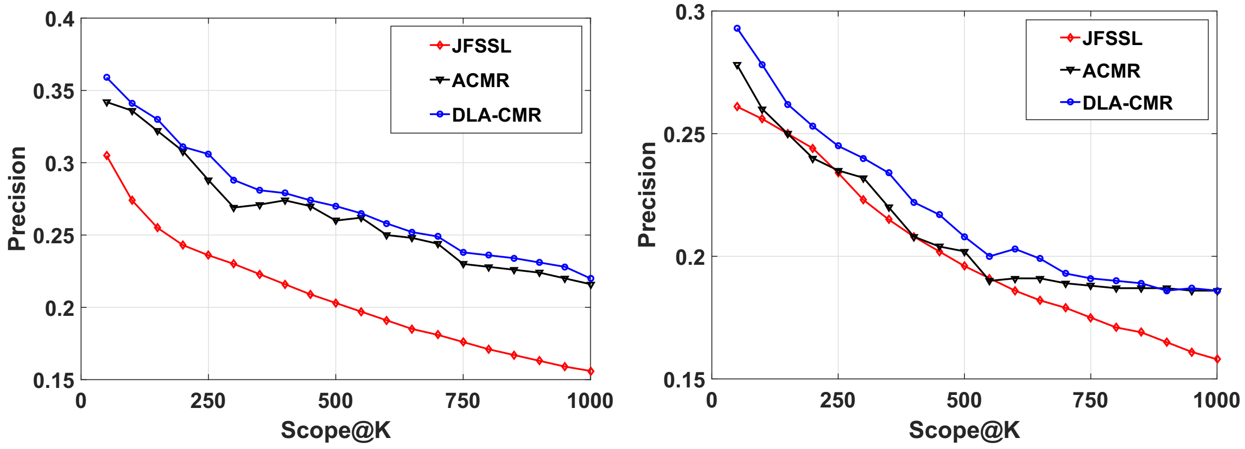
(c) T2I on Wikipedia-CNN dataset

(d) T2I on Pascal Sentence dataset

(e) Average MAP on Wikipedia-CNN dataset

(f) Average MAP on Pascal Sentence dataset

Fig. 4. MAP performance of each class on Wikipedia-CNN dataset and Pascal Sentence dataset.



(a) Txt2img

(b) Img2txt

Fig. 5. Precision-scope curves on the Wikipedia dataset with K ranges from 50 to 1000.

Table 4
MAP comparison between DLA-CMR and its variants on Wikipedia, Wikipedia-CNN and Pascal Sentence dataset. AL refers to adversarial learning.

Datasets	Methods	MAP scores		
		I2I	T2I	Average
Wikipedia	DLA-CMR without AL	0.2364	0.2198	0.2281
	DLA-CMR without tr	0.3217	0.2468	0.2843
	DLA-CMR without fd	0.1311	0.1274	0.1293
	full DLA-CMR	**0.369**	**0.261**	**0.315**
Wikipedia-CNN	DLA-CMR without AL	0.4081	0.3953	0.4017
	DLA-CMR without tr	0.5061	0.4017	0.4539
	DLA-CMR without fd	0.2317	0.2281	0.2299
	full DLA-CMR	**0.539**	**0.453**	**0.496**
Pascal Sentence	DLA-CMR without AL	0.4716	0.4806	0.4761
	DLA-CMR without tr	0.4832	0.5210	0.5021
	DLA-CMR without fd	0.2671	0.2834	0.2753
	full DLA-CMR	**0.498**	**0.549**	**0.522**

Table 5
MAP comparison between DLA-CMR and its variant on Wikipedia, Wikipedia-CNN and Pascal Sentence dataset. DL refers to dictionary learning.

Datasets	Methods	MAP scores		
		I2T	T2I	Average
Wikipedia	DLA-CMR without DL	0.349	0.248	0.299
	full DLA-CMR	**0.369**	**0.261**	**0.315**
Wikipedia-CNN	DLA-CMR without DL	0.508	0.423	0.466
	full DLA-CMR	**0.539**	**0.453**	**0.496**
Pascal Sentence	DLA-CMR without DL	0.472	0.506	0.489
	full DLA-CMR	**0.498**	**0.546**	**0.522**

Table 6
The running time of DL (dictionary learning), AL (adversarial learning) and their sum on four datasets.

Datasets	Modules	Running time(s)
Wikipedia	DL	219.68
	AL	143.61
	Sum	363.29
Wikipedia-CNN	DL	295.64
	AL	282.39
	Sum	578.03
Pascal Sentence	DL	257.38
	AL	96.12
	Sum	353.5
INRIA-Websearch	DL	6123.01
	AL	5283.69
	Sum	11406.7

modality classifier, which achieves the worst results in all of the variants. It is demonstrated that feature discrimination contributes more to performance than triplet ranking.

(2) Performance of dictionary learning model: As mentioned earlier, the dictionary learning model has a significant impact on experimental results. To demonstrate this conclusion, we conduct a variant and calculate their MAP scores as shown in Table 5. We can see that DLA-CMR achieves better performance than “DLA-CMR without DL”. It is demonstrated that dictionary learning can select significant features and improve accuracy.

As shown in Fig. 6, we show two examples of image query text and text query image on Wikipedia-CNN dataset. It can be seen that all results belong to the same category as the query instance, which verifies that DLA-CMR achieves better performance than compared methods.

In order to evaluate the time complexity of our DLA-CMR approach, several experiments are implemented on Intel(R) Xeon(R) E5-1650 v4 CPU 3.60 GHz × 12 machine with 32 GB RAM, GeForce GTX 1080 Ti/PCIe/SSE2, and the results are shown in Table 6. We can see that a large scale of dataset is more time consuming than a small scale dataset.

We carry out two experiments to verify the convergence of our proposed DLA-CMR approach. We show the convergence curves on Wikipedia-CNN dataset and Pascal Sentence dataset respectively as shown in Fig. 7. It can be observed that the total loss reduces as the learning epoch increases. Our DLA-CMR approach converges within about 120 epoch and 40 epoch respectively on these two datasets. These experimental results indicate that the convergence of DLA-CMR approach can be guaranteed.

5. Conclusion

In this paper, we have proposed an Adversarial Cross-Modal Retrieval Based on Dictionary Learning Algorithm (DLA-CMR) which can obtain discriminative feature representations for cross-media retrieval. On the one hand, we make full use of dictionary learnings ability of feature reconstruction to utilize reconstruction coefficients instead of raw features. On the other hand, we integrate triplet ranking, feature discrimination and modality classifier into an adversarial model. Transformed features are statistically indis-

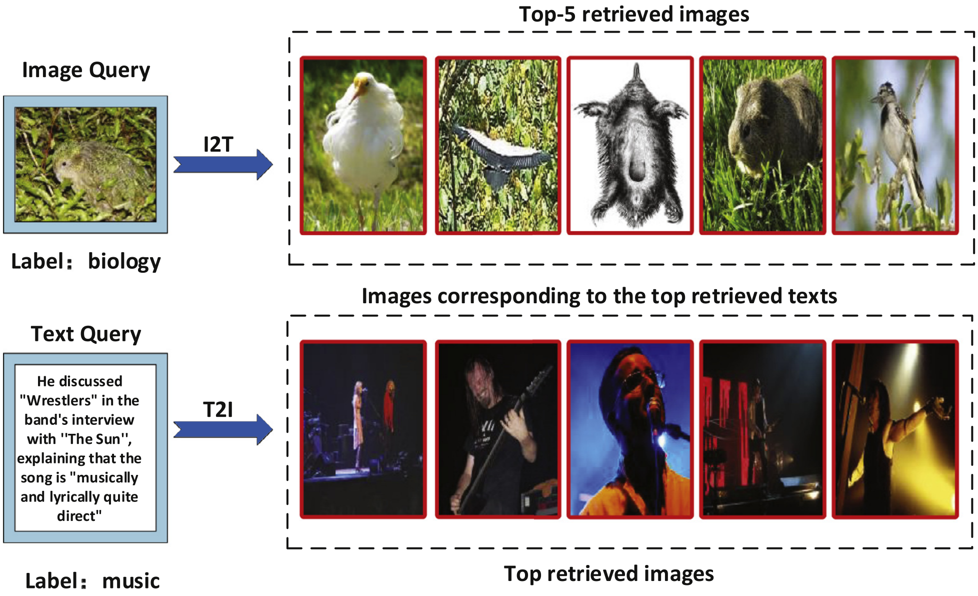

Top-5 retrieved images
Image Query
I2T
Label：biology
Images corresponding to the top retrieved texts
Text Query
He discussed "Wrestlers" in the band's interview with "The Sun", explaining that the song is "musically and lyrically quite direct"
T2I
Label：music
Top retrieved images

Fig. 6. Two examples of image query text and text query image on Wikipedia-CNN dataset. For the example of image query text, we use the corresponding images of retrieved texts to demonstrate the results. The red border images represent the correct retrieval.

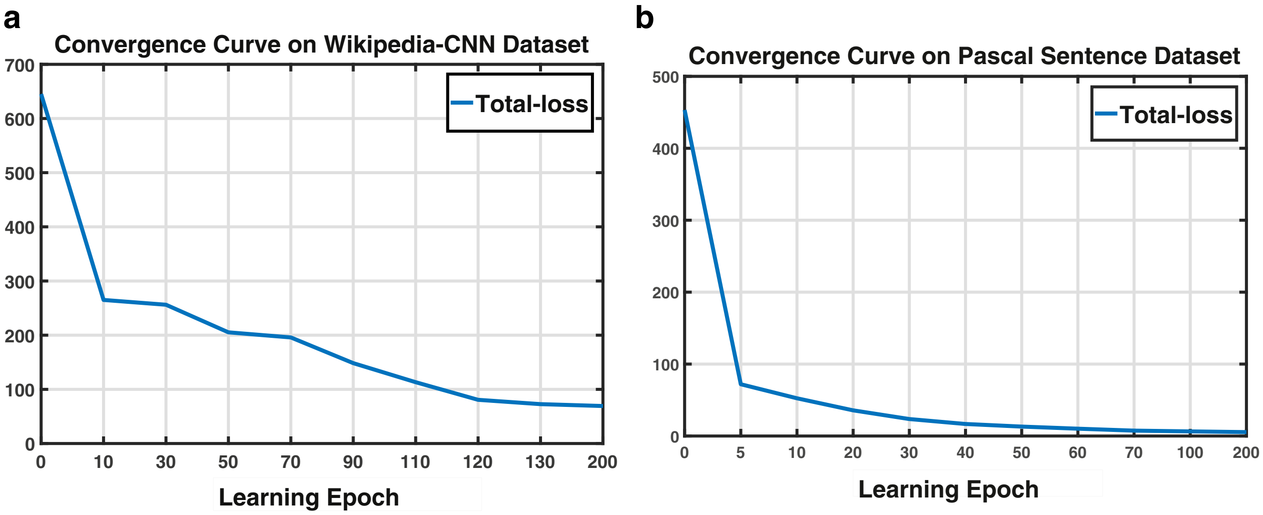


Fig. 7. The total loss varies from learning epoch, as computed for DLA-CMR on (a) Wikipedia-CNN and (b) Pascal Sentence dataset.

tinguishable and highly correlated in common space. Comprehensive experimental results on 4 widely-used datasets verify the good performance of our DLA-CMR method.

Conflict of interest

None.

Acknowledgment

The work is partially supported by the National Natural Science Foundation of China (Nos. 61572298, 61772322, U1836216, 61802236, 61601268) and the Technology and Development Project of Shandong (Nos. 2017GGX10117 and 2017CXGC0703).

References

[1] Meng, Zhao, Huaxiang, Zhang, Lili, Meng, An angle structure descriptor for image retrieval, China Commun. 13 (8) (2016) 222–230.
[2] C. Yan, H. Xie, J. Chen, Z.J. Zha, X. Hao, Y. Zhang, Q. Dai, A fast Uyghur text detector for complex background images, IEEE Trans. Multimed. 20 (12) (2018) 3389–3398.
[3] C. Yan, H. Xie, S. Liu, Y. Jian, Q. Dai, Effective uyghur language text detection in complex background images for traffic prompt identification, IEEE Trans. Intell. Trans. Syst. 19 (1) (2018) 220–229.
[4] H. Fan, Z. Xu, L. Zhu, C. Yan, J. Ge, Y. Yang, Watching a small portion could be as good as watching all: Towards efficient video classification, in: IJCAI, Vol. 2, 2018, pp. 705–711.
[5] Y.X. Peng, W.W. Zhu, Y. Zhao, C.S. Xu, Q.M. Huang, H.Q. Lu, Q.H. Zheng, T.J. Huang, W. Gao, Cross-media analysis and reasoning: advances and directions, Front. Inf. Technol. Electron. Eng. 18 (1) (2017) 44–57.
[6] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2372–2385.
[7] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, Q. Dai, Cross-modality bridging and knowledge transferring for image understanding, IEEE Trans. Multimed. (2019), doi:10.1109/TMM.2019.2903448.
[8] C. Wang, H. Yang, C. Meinel, Deep semantic mapping for cross-modal retrieval, in: Proceedings of the IEEE International Conference on TOOLS with Artificial Intelligence, 2016, pp. 234–241.
[9] 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.
[10] T. Yao, X. Kong, H. Fu, Q. Tian, Semantic consistency hashing for cross-modal retrieval, Neurocomputing 193 (C) (2016) 250–259.
[11] F. Feng, R. Li, X. Wang, Deep correspondence restricted Boltzmann machine for cross-modal retrieval, Neurocomputing 154 (C) (2015) 50–60.

[12] Q.Y. Jiang, W.J. Li, Deep cross-modal hashing, in: Computer Vision and Pattern Recognition, 2017, pp. 3270–3278.

[13] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, IEEE Trans. Circuits Syst. Video Technol. 24 (6) (2014) 965–978.

[14] Y. Peng, X. Zhai, Y. Zhao, X. Huang, Semi-supervised cross-media feature learning with unified patch graph regularization, IEEE Trans. Circuits & Systems for Video Technology 26 (3) (2016) 583–596.

[15] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, W. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2013.

[16] S. Bahrampour, N. Nasrabadi, A. Ray, W. Jenkins, Multimodal task-driven dictionary learning for image classification, IEEE Trans. Image Process. 25 (1) (2015) 24–38.

[17] F. Shang, H. Zhang, J. Sun, L. Liu, H. Zeng, A cross-media retrieval algorithm based on consistency preserving of collaborative representation, J. Adv. Comput. Intell. Intell. Informat. 22 (2) (2018) 280–289.

[18] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the International Conference on International Conference on Machine Learning, 2013, pp. III–1247.

[19] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the International Conference on Neural Information Processing Systems, 2014, pp. 2672–2680.

[20] X. Zhao, G. Ding, Y. Guo, J. Han, Y. Gao, Tuch: Turning cross-view hashing into single-view hashing via generative adversarial nets, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 3511–3517.

[21] X. Mao, S. Wang, L. Zheng, Q. Huang, Semantic invariant cross-domain image generation with generative adversarial networks, Neurocomputing (2018) 55–63.

[22] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, ICML 48 (2016) 1060–1069. arXiv preprint arXiv: 1605.05396.

[23] M. Mirza, S. Osindero, Conditional generative adversarial nets, Comput. Sci. (2014) 2672–2680. arXiv preprint arXiv: 1411.1784.

[24] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, Comput. Sci. (2015) arXiv preprint arXiv:1511.06434.

[25] J. Zhang, Y. Peng, M. Yuan, Unsupervised generative adversarial cross-modal hashing, in: Proceedings of the AAAI, 2017, pp. 539–546.

[26] J. Zhang, Y. Peng, M. Yuan, Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network, IEEE Trans. Cybern. (2018), doi:10.1109/TCYB.2018.2868826.

[27] Y. Peng, J. Qi, Y. Yuan, CM-GANs: Cross-modal generative adversarial networks for common representation learning, ACM Trans. Multimed. Comput. Commun. Appl. 15 (1) (2019) 22:1–22:24.

[28] 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.

[29] X. Xu, Y. Yang, A. Shimada, R.I. Taniguchi, L. He, Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts, in: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 847–850.

[30] Z. Zhu, Y. Chai, H. Yin, Y. Li, Z. Liu, A novel dictionary learning approach for multi-modality medical image fusion, Neurocomputing 214 (2016) 471–482.

[31] X. Xu, L. He, H. Lu, L. Gao, Y. Ji, Deep adversarial metric learning for cross–modal retrieval, World Wide Web 22 (2) (2019) 657–672.

[32] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, H.T. Shen, Unsupervised cross-modal retrieval through adversarial learning, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2017, pp. 1153–1158.

[33] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, in: Proceedings of the ICCV, 2015, pp. 2623–2631.

[34] A. Karpathy, F.F. Li, Deep visual-semantic alignments for generating image descriptions, in: Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[35] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, S. Yan, Modality-dependent cross-media retrieval, ACM Trans. Intell. Syst. Technol. 7 (4) (2016) 1–13.

[36] 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.

[37] 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.

[38] 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, 2013, pp. 2088–2095.

[39] J. Wu, Z. Lin, H. Zha, Joint latent subspace learning and regression for cross–modal retrieval, in: Proceedings of the International ACM SIGIR Conference, 2017, pp. 917–920.

[40] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Trans. Multimed. PP (99) (2017). 1–1

[41] B. Zhang, L. Zhu, J. Sun, H. Zhang, Cross-media retrieval with collective deep semantic learning, Multimed. Tools Appl. 77 (17) (2018) 22247–22266.

Fei Shang received her B.S. degree in computer science and technology from Shandong Normal University, China, in 2017. She is currently pursuing the master degree in the School of Information Science & Engineering from the same university. Her research interests include cross-modal retrieval, machine learning, and deep learning. She is a student member of the CCF.

Huaxiang Zhang received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2004. He was an Associated Professor with the Department of Computer Science, Shandong Normal University, Jinan, China, from 2004 to 2005, where he is currently a Professor with the School of Information Science and Engineering. He has authored over 100 journal and conference papers and has been granted eight invention patents. His current research interests include machine learning, pattern recognition, evolutionary computation, and Web information processing.

Lei Zhu received the B.S. degree (2009) at Wuhan University of Technology, the Ph.D. degree (2015) at Huazhong University of Science and Technology. He is currently a full Professor with the School of Information Science and Engineering, Shandong Normal University, China. He was a Research Fellow under the supervision of Prof. Heng Tao Shen at the University of Queensland (2016–2017), and Dr. Jialie Shen at the Singapore Management University (2015–2016). His research interests are in the area of large-scale multimedia content analysis and retrieval.

Jiande Sun received the Ph.D. degree in communication and information system from Shandong University, Jinan, China, in 2005. He has been the visiting researcher in Technical University of Berlin, University of Konstanz, and Carnegie Mellon University, and a Post-Doctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and with the State Key Laboratory of Digital-Media Technology, Hisense Group. He has published more than 60 journal and conference papers. He is the co-author of two books.