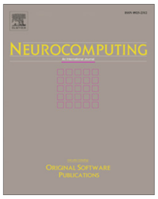
Neurocomputing 509 (2022) 85–93

Contents lists available at ScienceDirect

Neurocomputing

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



Cross-media retrieval of scientific and technological information based on multi-feature fusion

Yang Jiang, Junping Du *, Zhe Xue, Ang Li

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications Beijing 100876, China

ARTICLE INFO

Article history:
Received 5 July 2021
Revised 24 April 2022
Accepted 18 June 2022
Available online 8 July 2022
Communicated by Zidong Wang

Keywords:
Cross-media retrieval
Adversarial learning
Neural network

ABSTRACT

In the era of big data, People's lives are filled with all kinds of information. Scientific and technological information is utilized for scholars to understand the current technology trends, and to think about the source of information for future development prospects. More and more scholars are no longer satisfied with single-modal retrieval methods. However, to get more intelligent cross-media retrieval results we should give higher requirements to the search engine. And how to span the semantic gap between different modalities is a key issue that needs to be solved. In response to the above problems, this paper proposes a Multi-feature Fusion based Cross-Media Retrieval (MFCMR) method. Our method is capable of integrating multiple features to promote semantic understanding, and adopting adversarial learning to further improve the accuracy of public subspace representation. Then we use similarity in the same space to sort the retrieval results. We conduct a lot of experiments on real datasets, and the results show that our method obtains better cross-media retrieval performance than other methods.

1. Introduction

With the rapid development of the Internet, various data on the network increases rapidly [1,2]. These massive scientific and technological repositories take the network as the carrier, and the storage and dissemination of knowledge effectively promote the rapid development and progress of science and technology. At present, the analysis of scientific and technological resources of single media such as paper, and scholars has been well developed, however, the analysis of cross-media scientific and technological information data has not been well studied. Scientific and technological information has its unique characteristics, for it contains not only text data, but also image data. To obtain scientific and technological information from multiple modalities, make scholars more fully understand the current research trend, and promote the development of production and research, developing a new cross-media retrieval algorithm for scientific and technological information is of great significance for researchers to better grasp the current scientific research situation.

The difficulty of cross-media retrieval lies in how to map heterogeneous modalities to a common subspace. There are two common space learning methods, namely the traditional direct mapping method [3] and the deep learning method [4]. However, the traditional direct mapping method has a simple structure and cannot understand high-dimensional cross-modal semantics deeply. With the development of deep network models, the use of multi-layer convolutional neural networks has become a hot spot for feature extraction and public space mapping. Andrew et al.[5] put forward a deep canonical correlation analysis method (DCCA) based on deep learning. However, for cross-media scientific and technological information data, there are still problems such as insufficient semantic understanding and poor subspace mapping effects that have not been solved.

Therefore, this paper proposes a Multi-feature Fusion based Cross-Media Retrieval (MFCMR). This method mainly uses an adversarial learning approach to train two neural network models, namely the feature mapping network and the modality discriminant network. Feature mapping network, which acts as a generator for adversarial learning to map features from text, image, temporal, and author to the same semantic space, we use semantic similarity to train the feature mapping network. The modality discriminant network acts as a discriminator for adversarial learning to distinguish the original modalities of the data mapped to the same semantic space, and trains the modality discriminant network with the real modalities of the data. Then, after projecting the text or image into the same semantic space through the feature mapping network, the search results are obtained according to the distance from other data in this space.

* Corresponding author.
E-mail address: junpingdu@126.com (J. Du).

https://doi.org/10.1016/j.neucom.2022.06.061
0925-2312/© 2022 Published by Elsevier B.V.

The MFCMR method is intended to solve the problem of cross-media search for scientific and technological information. The contributions of this article are as follows:

(1) We integrate multiple features of scientific and technological information into the training process, which can improve the semantic understanding ability to a certain extent, and use the adversarial learning method to further enhance the accuracy of the subspace model.
(2) We design an embedding loss function combining modality loss and consistency regularization loss in the feature mapping network, which can help to eliminate the difference between different modalities of data with the same semantics.
(3) We use the cross-media scientific and technological information data to verify the MFCMR method. The results show that MFCMR can effectively improve the accuracy of cross-media retrieval of scientific and technological information compared to state-of-the-art algorithms.

2. Related Work

2.1. Cross-media Data Semantic Learning

In the research on the semantic representation of cross-media scientific and technological text data, the analysis and mining of semantic features must be combined with machine learning and data mining technology to accurately obtain the subject, field, interest,and other tags of the technology entity. The traditional topic models [6–8] use matrix decomposition technology to map the original terms to a low-dimensional semantic space, which can better express semantic features; pLSA [9] probabilizes the LSA and gives the topic model a probabilistic meaning; Latent Dirichlet Allocation (LDA) [10,11] combines the pLSA model with Bayesian modeling to form a three-layer Bayesian model for documents, topics, and terms. By assuming that the document is composed of a mixed distribution of topics, each topic obeys a polynomial distribution, and Dirichlet is introduced as a priori information. These methods have achieved excellent results in text mining and information retrieval. Combined with the use of multiple machine learning algorithms, Yang et al. [12] mainly considered the attributes of scholars and papers, and constructed a complete analysis model of scholar user portraits. Based on the topic model, it mines the latent semantic information of the text, extracts the scholars' heterogeneous network connection structure information, and forms the scholars' interest tags. Li et al. [13] improved the distribution similarity function based on the topic distribution recognition of the text based on the LDA topic model, carried out hierarchical clustering of the courses, and found the correlation between the topic texts based on different levels of abstraction. Pang et al. [14] put forward a microblogging generation method MRT-LDA which takes the relations between Chinese micro-blog documents and other Chinese micro-blog documents into consideration to help topic mining in micro-blog.

The semantic representation of cross-media scientific and technological image data has drawn considerable research interest [15,16]. Simonyan et al. [17] proposed the VGGNet model. The relationship between the depth of the convolutional neural network and its performance is explored. Through repeated stacking of 3*3 small convolution kernels and 2*2 maximum pooling layers, a volume of 16 to 19 layers deep neural network is successfully constructed. Dhankhar et al. [18] used the combination of ResNet-50 and VGG-16 to recognize facial expressions, and obtained better results in the KDEF data set. Qassim et al. [19] proposed a compression convolutional neural network model called Residual Squeeze VGG-16 to solve the velocity and scale problems. The model size of this algorithm is small and the training speed is fast. The method uses residual learning to make convergence faster and generalization better, and solves the degradation problem. Yuan et al. [20] proposed a low-rank matrix regression method for feature extraction and feature selection. This algorithm can make full use of image structure information and improve the accuracy of image feature extraction. Peng et al. [21] proposed a strategy of aggregating low-order CNN feature maps to generate local features, which can effectively solve the shortcoming that global CNN features cannot effectively describe local details and abstract more accurate image feature vectors.

2.2. Adversarial Learning Method

GANs [22,23] is a deep learning method that generates a model through adversarial learning, and generates a distribution which is close to the target distribution through neural network learning [24,25]. GANs are composed of generators and discriminators. The generator generates real data distribution through learning sample data as much as possible; the discriminator distinguishes whether the data source is real or the data generated by the generator, and influences the learning of the generator through the discrimination results. In the training process, when the discriminator cannot distinguish whether the data is real data or data generated by the generator, the generator is considered to achieve the best effect.

Researchers apply ideas to cross-media retrieval through research on adversarial learning. Wang et al. [26] proposed the adversarial cross-media retrieval (ACMR) method. The core idea is to seek to map the two modalities to a common subspace through the interaction between feature projector and the modality classifier. The feature mapper attempts to form an invariant representation of the modality in the common subspace to confuse the modality classifier. The modality classifier is composed of label prediction and triple constraint, which can minimize the distance of the same semantic vector in different modalities. He et al. [27] proposed a new unsupervised cross-modal retrieval method UCAL based on adversarial learning, which has better results for cross-media data with less annotations. Peng et al. [28] proposed a cross-modal structure to simulate the joint distribution of different modal data by generating adversarial network GANs. At the same time, they proposed a cross-modal convolutional autoencoder with weight sharing constraints to form a generative model.

Chun et al. [29] proposed a Probabilistic Cross-Modal Embedding (PCME) that can efficiently identify functional defects and obtain public representations of modal models in the embedding space. Messina et al. [30] proposed a Transformer Encoder Reasoning and Alignment Network (TERAN) which enforces fine-grained matching between images and sentences base components to preserve the informative richness of both modalities. Wang et al. [31] proposed a new supervised cross-modal hashing method, namely scalaBle Asymmetric discreTe Cross-modal Hashing (BATCH). This algorithm utilizes collective matrix factorization to learn labels and a common latent space of different modalities, and embeds labels into binary codes by minimizing the distance-distance difference problem.

3. MFCMR

3.1. Description of the MFCMR method

Cross-media scientific and technological information data has the following characteristics: It contains rich text and image data; the same author often publishes similar scientific and technological information; data with similar semantics will appear in similar periods, and so on. The general framework of the proposed MFCMR

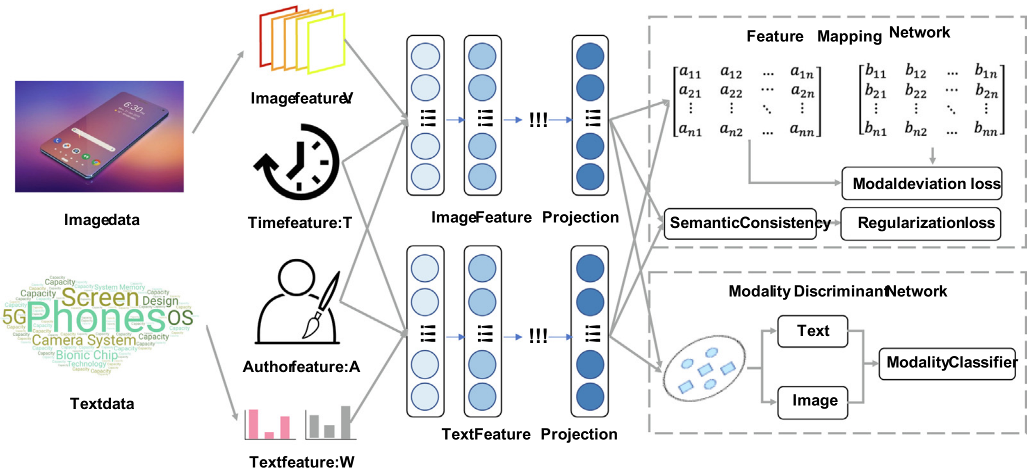

Imagedata
Textdata
Imagefeature:V
Timefeature:T
Authorfeature:A
Textfeature:W
ImageFeature Projection
TextFeature Projection
!!!
Feature Mapping Network
$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$
$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$
Modaldeviation loss
SemanticConsistency
Regularizationloss
Modality DiscriminantNetwork
Text
Image
ModalityClassifier
5G Phones OS Screen Design Capacity Camera System Bionic Chip Technology System Memory

Fig. 1. The general flowchart of the proposed MFCMR method.

method is shown in Fig. 1, the MFCMR method consists of a feature mapping network and a modal discrimination network. The ultimate goal is to map the vectors of different modalities to a common subspace.

The input of the feature mapping network is a multi-feature fusion vector, including an image feature vector, text feature vector, time feature vector,and author feature vector. The image mapping vector is obtained by the VGG16 neural network, the text mapping vector is obtained by the LDA topic model, and the time feature and author feature are obtained using a one-hot vector. The feature mapper is composed of a multi-feature semantic analysis network, which uses the output result of softmax as the semantic distribution mapped to the public space so that different modalities with the same semantics are close to each other, and different modalities with different semantics are far away from each other.

Denote $\boldsymbol{O} = \{\boldsymbol{o}_i = [\boldsymbol{v}_i, \boldsymbol{w}_i, \boldsymbol{t}_i, \boldsymbol{a}_i, \boldsymbol{l}_i]\}_{i=1}^{n}$,with $\boldsymbol{v}_i \in \boldsymbol{R}^{d_v}$ representing the d_v dimensional image feature,$\boldsymbol{w}_i \in \boldsymbol{R}^{d_w}$ denoting the d_w dimensional text feature vector, $\boldsymbol{t}_i \in \boldsymbol{R}^{d_t}$ representing the d_t dimensional time, $\boldsymbol{a}_i \in \boldsymbol{R}^{d_a}$ being the d_a dimensional author feature vector, and $\boldsymbol{l}_i \in \boldsymbol{R}^{d_l}$ representing the d_l dimensional label feature vector.We denote the image feature matrix, text feature matrix, time feature matrix, author feature matrix and label matrix for all instances in $\boldsymbol{O}$ as $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_n\} \in \boldsymbol{R}^{d_v * n}, \boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \ldots, \boldsymbol{w}_n\} \in \boldsymbol{R}^{d_w} * n, \boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \ldots, \boldsymbol{t}_n\} \in \boldsymbol{R}^{d_t * n}, \boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \ldots, \boldsymbol{a}_n\} \in \boldsymbol{R}^{d_a * n}$, and $\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3, \ldots, \boldsymbol{l}_n\} \in \boldsymbol{R}^{d_l * n}$, respectively. Cross-modal retrieval is to map the vectors of two modalities to a common subspace S, and the common subspace S is obtained by two functions: $\boldsymbol{S}_w = f_W(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{A}, \theta_w), \boldsymbol{S}_v = f_V(\boldsymbol{V}, \boldsymbol{T}, \boldsymbol{A}, \theta_v)$. f_W is a feature mapping function that considers text feature, time feature, and author feature, while f_V is a feature mapping function that considers image feature, time feature, and author feature. $\boldsymbol{S}_w \in \boldsymbol{R}^{p*n}$ and $\boldsymbol{S}_v \in \boldsymbol{R}^{p*n}$ are text and image feature in common subspace.

3.2. Feature Mapping Network

The objective function of the MFCMR is composed of the embedding loss function $L_{emb}(\theta_V, \theta_W, \theta_{imd})$ and the adversarial loss function $L_{adv}(\theta_D)$. The embedding loss function includes the label prediction loss function L_{imd}, the modality deviation loss function L_{mod} and the semantic consistency regularization loss function L_{reg} , such as the formula: $L_{emb}(\theta_V, \theta_W, \theta_{imd}) = L_{imd} + \alpha \cdot L_{mod} + \beta \cdot L_{reg}$,where α, β are all hyperparameters, the purpose is to balance the importance of different parts.

We obtain the label prediction loss function $L_{imd}(\theta_{imd})$ as:

$$L_{imd}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot (logp_i(\boldsymbol{v}_i) + logp_i(\boldsymbol{w}_i))) \tag{1}$$

where n is the number of instances within each mini-batch, y_i is the groundtruth of each instance, while $p_i(\cdot)$ is the probability distribution of image or text generation in instance $\boldsymbol{o}_i$, and θ_{imd} is the classifier parameter.

In order to ensure that data with the same semantics is close in different modalities, and data with different semantics is far away, a multi-feature semantic analysis network is designed. On the basis of calculating the similarity of data semantic distribution, the original data semantic distribution $l_{1\ldots n}$ is used to construct the semantic similarity distribution matrix $\boldsymbol{Sim}_L \in \boldsymbol{R}^{n*n}$. Then the semantic distribution of any two data is defined as $\boldsymbol{l}_x, \boldsymbol{l}_y$. We estimate their similarity $sim(\boldsymbol{l}_x, \boldsymbol{l}_y)$ as:

$$sim(\boldsymbol{l}_x, \boldsymbol{l}_y) = \frac{\boldsymbol{l}_{xi} \cdot \boldsymbol{l}_{yi}}{\|\boldsymbol{l}_{xi}\| \cdot \|\boldsymbol{l}_{yi}\|} = \frac{\sum_{i=1}^{d_l} \boldsymbol{l}_{xi} \cdot \boldsymbol{l}_{yi}}{\sqrt{\sum_{i=1}^{d_l} (\boldsymbol{l}_{xi})^2} \cdot \sqrt{\sum_{i=1}^{d_l} (\boldsymbol{l}_{yi})^2}} \tag{2}$$

Here we need to calculate the similarity of all data.

$$\boldsymbol{Sim}_L(i,j) = sim(\boldsymbol{l}_i, \boldsymbol{l}_j) \tag{3}$$

After the semantic similarity matrix is obtained, the similarity matrix of the mapped data features is calculated. For any two sets of data $\boldsymbol{o}_x, \boldsymbol{o}_y$, extract the text feature, time feature, and author feature in $\boldsymbol{o}_x$, and get $\boldsymbol{S}_x$ after feature mapping; extract the text feature, time feature, and author feature in $\boldsymbol{o}_y$, and get $\boldsymbol{S}_y$ after feature mapping. Use the above formula to calculate $sim(\boldsymbol{S}_x, \boldsymbol{S}_y)$, and perform similarity calculation on all data after feature mapping so as to obtain the mapped data feature similarity matrix $\boldsymbol{Sim}_s \in \boldsymbol{R}^{n*n}$.

$$\begin{aligned} \boldsymbol{Sim}_S(\boldsymbol{i}, \boldsymbol{j}) &= sim(\boldsymbol{s}_{vi}, \boldsymbol{s}_{wj}) \\ &= sim(f_V(\boldsymbol{v}_i, \boldsymbol{t}_i, \boldsymbol{a}_i, \theta_V), f_W(\boldsymbol{w}_j, \boldsymbol{t}_j, \boldsymbol{a}_j, \theta_W)) \end{aligned} \tag{4}$$

We chooses ℓ_2 norm to measure the difference between two similarity matrices, and defines the difference value as the modality deviation value $L_{mod}(\theta_V, \theta_W)$.

$$\begin{aligned} L_{mod}(\theta_V,\theta_W) &= \ell_2(\boldsymbol{Sim}_L,\boldsymbol{Sim}_S) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left\| sim(\boldsymbol{l}_i,\boldsymbol{l}_j) - sim(f_V(\boldsymbol{v}_i,\boldsymbol{t}_i,\boldsymbol{a}_i,\theta_V), f_w(\boldsymbol{w}_j,\boldsymbol{t}_j,\boldsymbol{a}_j,\theta_W))\right\|_2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left\| \frac{\sum_{k=1}^{d_l}\boldsymbol{l}_{ik}\cdot\boldsymbol{l}_{jk}}{\sqrt{\sum_{k=1}^{d_l}(\boldsymbol{l}_{ik})^2}\cdot\sqrt{\sum_{k=1}^{d_l}(\boldsymbol{l}_{jk})^2}} - \frac{\sum_{k=1}^{m} f_V(\boldsymbol{v}_i,\boldsymbol{t}_i,\boldsymbol{a}_i,\theta_V)_k \cdot f_w(\boldsymbol{w}_j,\boldsymbol{t}_j,\boldsymbol{a}_j,\theta_W)_k}{\sqrt{\sum_{k=1}^{m}(f_V(\boldsymbol{v}_i,\boldsymbol{t}_i,\boldsymbol{a}_i,\theta_V)_k)^2}\cdot\sqrt{\sum_{k=1}^{d_l}(f_W(\boldsymbol{w}_j,\boldsymbol{t}_j,\boldsymbol{a}_j,\theta_W)_k)^2}} \right\|_2 \end{aligned} \tag{5}$$

In order to ensure that the samples with the same semantics are constantly close during the mapping process, the semantic consistency regularization loss function L_{reg} is defined. To get the distribution of image feature $\boldsymbol{V}$ and text feature $\boldsymbol{W}$, we calculate their set of class centers $\boldsymbol{C}_V = \{\boldsymbol{c}_1^v, \boldsymbol{c}_2^v, \ldots, \boldsymbol{c}_n^v\}$ and $\boldsymbol{C}_W = \{\boldsymbol{c}_1^w, \boldsymbol{c}_2^w, \ldots, \boldsymbol{c}_n^w\}$. We calculate the class center of $\boldsymbol{c}_{1\ldots k}^v$ and $\boldsymbol{c}_{1\ldots k}^w$ by:

$$\boldsymbol{c}_j^v = \frac{1}{n_j}\sum_{i=1}^{n_j}\boldsymbol{v}_j^i \tag{6}$$

$$\boldsymbol{c}_j^w = \frac{1}{n_j}\sum_{i=1}^{n_j}\boldsymbol{w}_j^i \tag{7}$$

We define d_1 to minimize the intra-class distance in the same modality.

$$d_1 = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left(\left\|\boldsymbol{c}_j^v - \boldsymbol{v}_j^i\right\|_2 + \left\|\boldsymbol{c}_j^w - \boldsymbol{w}_j^i\right\|_2\right)\right) \tag{8}$$

For different modalities, we define d_2 to minimize the distance between the center of the class.

$$d_2 = \frac{1}{k}\sum_{j=1}^{k}\left\|\boldsymbol{c}_j^w - \boldsymbol{c}_j^v\right\|_2 \tag{9}$$

For different modalities, we define d_3, d_4 to minimize the distance between the class center of the minimized modality and another modality sample with the same semantics.

$$d_3 = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left\|\boldsymbol{c}_j^v - \boldsymbol{w}_j^i\right\|_2\right) \tag{10}$$

$$d_4 = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left\|\boldsymbol{c}_j^w - \boldsymbol{v}_j^i\right\|_2\right) \tag{11}$$

Therefore, the semantic consistency regularization can be expressed as:

$$L_{reg} = (d_1 + d_2 + d_3 + d_4)/4 \tag{12}$$

3.3. Modality Discriminant Network

The modality discrimination network is used to distinguish the modalities of the data mapped to the common subspace, defining the data label after image mapping as 0, and the data label after text mapping as 1. The modality discrimination network correctly discriminates the original modality of the mapped data as much as possible and defines the loss function of this network as the deviation value of the modality prediction. The adversarial loss function is defined as:

$$\begin{aligned} L_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\big(&logD(f_W(\boldsymbol{w}_j,\boldsymbol{t}_j,\boldsymbol{a}_j,\theta_W),\theta_D) \\ &+log(1 - D(f_V(\boldsymbol{v}_i,\boldsymbol{t}_i,\boldsymbol{a}_i,\theta_V),\theta_D))\big) \end{aligned} \tag{13}$$

where $D(\boldsymbol{x},\theta_D)$ indicates that the network determines the probability $\boldsymbol{x}$ is text.

We obtain the objective loss function L_{loss} as:

$$L_{loss} = L_{emb}(\theta_V,\theta_W,\theta_{imd}) - L_{adv}(\theta_D) \tag{14}$$

Then according to the adversarial learning mechanism, the optimal characteristics of the two processes are continuously trained.

$$(\theta_V,\theta_W,\theta_{imd}) = \arg\min\left(L_{emb}(\theta_V,\theta_W,\theta_{imd}) - L_{adv}(\theta_D)\right) \tag{15}$$

$$\theta_D = \arg\max\left(L_{emb}(\theta_V,\theta_W,\theta_{imd}) - L_{adv}(\theta_D)\right) \tag{16}$$

The ultimate goal of the optimization process of MFCMR is to train two mapping functions f_W and f_V. The first process determines θ_D in $L_{adv}(\theta_D)$, then take the minimum value for the loss function, and calculate the three parameter values $\theta_V, \theta_W, \theta_{imd}$ of $L_{emb}(\theta_V,\theta_W,\theta_{imd})$. The second process brings the $\theta_V, \theta_W, \theta_{imd}$ parameters calculated in the first process into the loss function to determine $L_{emb}(\theta_V,\theta_W,\theta_{imd})$, then maximize the loss function, to obtain parameter θ_D.

The MFCMR training process is shown in Algorithm 1.

Algorithm1: Pseudocode of optimizing the MFCMR method

```
Input:
Based on the idea of minibatch, extract the following feature matrix;
Parameters $k$,$\mu$,$\lambda$ and the number of samples in each batch of each mode $m$;
Randomly initialize the parameters of the method;
Output:
Feature mapping function:$f_V(V,M,A)$, $f_W(W,M,A)$;
whileNot Converge do
    while k >0 do
        Update optimization parameters through stochastic gradient descentc $\theta_V$, $\theta_W$, $\theta_{imd}$:
        $\theta_V \leftarrow \theta_V - \mu \cdot \nabla_{\theta_V}\frac{1}{m}(L_{emb} - L_{adv})$;
        $\theta_W \leftarrow \theta_W - \mu \cdot \nabla_{\theta_W}\frac{1}{m}(L_{emb} - L_{adv})$;
        $\theta_{imd} \leftarrow \theta_{imd} - \mu \cdot \nabla_{\theta_{imd}}\frac{1}{m}(L_{emb} - L_{adv})$;
        $k = k - 1$;
    end while
    Optimize parameters through gradient $\theta_D$;
    $\theta_D \leftarrow \theta_D + \mu \cdot \lambda \cdot \nabla_{\theta_D}\frac{1}{m}(L_{emb} - L_{adv})$;
end while
return $f_V(V,T,A)$ and $f_W(W,T,A)$;
```

Table 1
Statistics of the datasets in our experiments.

	Training Instance	Test Instance	Labels	Image Feature	Text Feature
Dataset 1	5872	2516	12	4096d VGG	50d LDA 3000d BoW
Dataset 2	6515	2792	12	4096d VGG	50d LDA 3000d BoW

Table 2
Comparison of cross-media retrieval performance on dataset 1.

	map@5			map@20			map@50		
	txt2img	img2txt	avg	txt2img	img2txt	avg	txt2img	img2txt	avg
CCA	0.2382	0.1782	0.2082	0.2393	0.1808	0.2101	0.2279	0.1761	0.2020
JFSSL	0.3859	0.2781	0.3320	0.3947	0.2874	0.3411	0.3692	0.2698	0.3195
Corr-AE	0.3490	0.3321	0.3406	0.3529	0.3377	0.3485	0.3210	0.3291	0.3251
JRL	0.4231	0.3804	0.4018	0.4417	0.3872	0.4145	0.4162	0.3979	0.4071
Deep-SM	0.4876	0.4620	0.4748	0.4936	0.4671	0.4803	0.4681	0.4592	0.4637
ACMR	0.4962	0.4452	0.4707	0.5119	0.4360	0.4740	0.4925	0.4273	0.4599
BATCH@16bit	0.4893	0.4460	0.4677	0.4968	0.4472	0.4720	0.4867	0.4284	0.4576
BATCH@32bit	0.4967	0.4536	0.4752	0.4985	0.4553	0.4769	0.4903	0.4289	0.4596
SSACR	0.4981	0.4537	0.4759	0.5072	0.4572	0.4822	0.5021	0.4341	0.4681
PAN	0.4994	0.4523	0.4759	0.5088	0.4583	0.4836	0.5064	0.4364	0.4714
OURs	**0.5192**	**0.4703**	**0.4948**	**0.5307**	**0.4602**	**0.4955**	**0.5157**	**0.4520**	**0.4839**

Table 3
Comparison of cross-media retrieval performance on dataset 2.

	map@5			map@20			map@50		
	txt2img	img2txt	avg	txt2img	img2txt	avg	txt2img	img2txt	avg
CCA	0.2271	0.1793	0.2032	0.2329	0.1865	0.4194	0.2184	0.1876	0.2030
JFSSL	0.3752	0.2707	0.3230	0.3852	0.2811	0.3332	0.3731	0.2684	0.3208
Corr-AE	0.3428	0.3273	0.3351	0.3379	0.3220	0.3300	0.3084	0.3126	0.3105
JRL	0.4262	0.4079	0.4171	0.4251	0.4157	0.4204	0.4077	0.3821	0.3949
Deep-SM	0.4685	0.4219	0.4452	0.4627	0.4163	0.4395	0.4302	0.4063	0.4183
ACMR	0.4849	0.4463	0.4656	0.4998	0.4384	0.4691	0.4712	0.4258	0.4485
BATCH@16bit	0.4867	0.4474	0.4671	0.5013	0.4496	0.4755	0.4727	0.4256	0.4492
BATCH@32bit	0.4924	0.4473	0.4699	0.5042	0.4524	0.4783	0.4733	0.4270	0.4502
SSACR	0.4920	0.4582	0.4751	0.5021	0.4565	0.4793	0.4731	0.4316	0.4524
PAN	0.4922	0.4585	0.4754	0.5037	0.4558	0.4798	0.4746	0.4325	0.4536
OURs	**0.5136**	**0.4670**	**0.4903**	**0.5221**	**0.4572**	**0.4897**	**0.4952**	**0.4501**	**0.4727**

4. Experiment

4.1. Dataset And Evaluation Metrics

The experiment uses real data collected from "Kuaikeji"[1] websites to verify the effectiveness of MFCMR. The scientific and technological cross-media information data is divided into two datasets according to the period time. Dataset 1 is defined as the period time 2017.01.01–2018.12.31, and dataset 2 is defined as the period time 2019.01.01–2020.12.31.

Dataset 1 and dataset 2 have a total of 8,388 pairs and 9,307 pairs of data. The dataset 1 contains ground truth clusters for 12 different labels, computer(739 clusters), automobile(852 clusters), science(984 clusters), evaluation(251 clusters), Android(931 clusters), Apple(823 clusters), Huawei(787 clusters), CPU(183 clusters), graphics(803 clusters), game(714 clusters), audio and video(692 clusters), astronavigation(629 clusters).The dataset 2 contains ground truth clusters for 12 different labels, computer(752 clusters), automobile(805 clusters), Android(882 clusters), Apple(862 clusters), Huawei(854 clusters), Mars(748 clusters), the earth(832 clusters), CPU(178 clusters), graphics(803 clusters), game(742 clusters), audio and video(853 clusters), COVID-19(867 clusters).70% of data is used as the training set, and 30% of data is used as the testing set. Data and characteristics are shown in Table 1.

We use mAP (mean Average Precision) to evaluate the retrieval results. mAP comprehensively considers ranking information and ranking accuracy, that is, the average value of the retrieval accuracy of each relevant document. The larger the mAP, the better the performance of the algorithm.

4.2. Baselines

We compare our MFCMR method with the following baselines:

Canonical correlation analysis (CCA). This method [3] learns a subspace to maximize the pairwise association between two sets of heterogeneous data.

Joint feature selection and subspace learning (JFSSL). This method [32] proposes an iterative algorithm to jointly solve two problems: the measure of relevance and coupled feature selection.

Correspondence autoencoder (Corr-AE). This method [33] is constructed by correlating hidden representations of two unimodal autoencoders.

Joint representation learning (JRL). This method [34] can explore jointly the correlation and semantic information in a unified optimization framework.

[1] https://news.mydrivers.com

Deep semantic matching (Deep-SM). This method [35] uses two different deep neural networks to map images and texts into a homogeneous semantic space.

Adversarial cross-modal retrieval (ACMR). This method[26] uses adversarial learning to further optimize the distance between the same modality and different modalities and uses the triple constraint method to reduce the distance of the same semantic data between different modalities.

ScalaBle Asymmetric discreTe Cross-modal Hashing (BATCH). This method [31] exploits collective matrix factorization to learn labels and a common latent space of different modalities, and embeds labels into binary codes by minimizing the distance-distance difference problem.

Semantic Similarity based Adversarial Cross Media Retrieval (SSACR). This method [36] uses adversarial learning to train two networks alternately and finally makes the data obtained from the feature mapping network consistent with the original data semantically.

Prototype-based Adaptive Network (PAN). This method [37] leverages a unified prototype to represent each semantic category across modalities, which provides discriminative information of different categories and takes unified prototypes as anchors to learn cross-modal representations adaptively.

MFCMR_N. MFCMR method uses only the image feature and text feature.

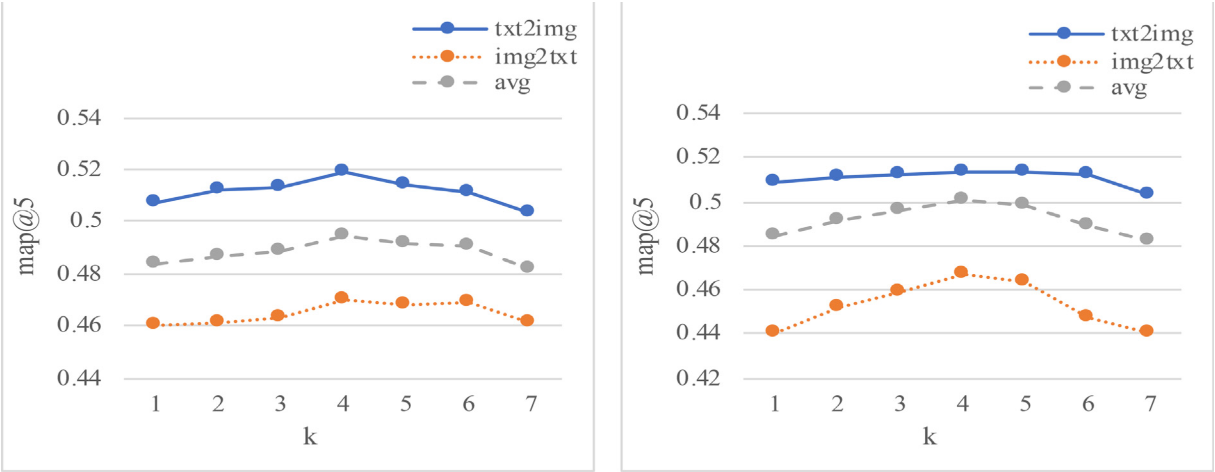
MFCMR_T. MFCMR method uses only the image feature, text feature, and time feature.

MFCMR_A: MFCMR method uses only the image feature, text feature, and author feature.

4.3. Experimental Analysis

We use mAP to measure the accuracy of the algorithm, and calculate the mAP value through the top 5, top 20, and top 50 search results of the two tasks of text search image and image search text on dataset 1 and dataset 2. The mAP values obtained by the comparison algorithm are shown in Table 2, and Table 3.

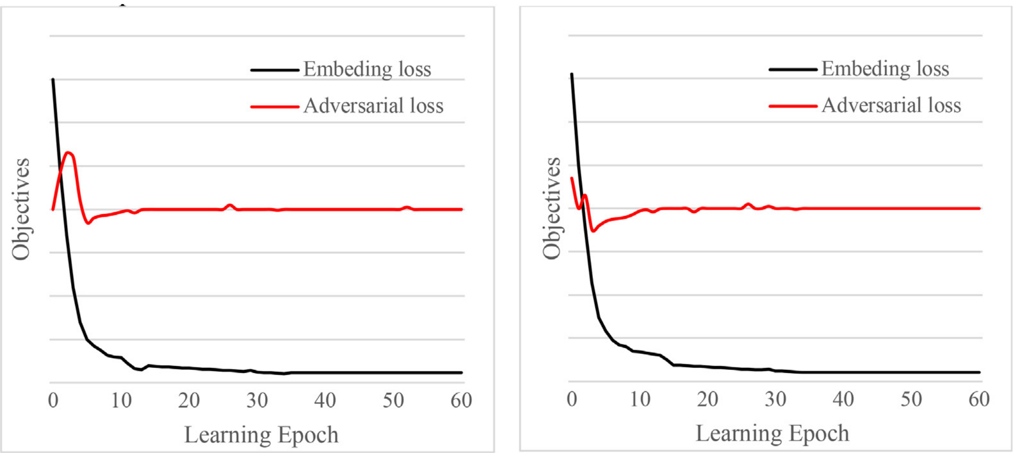
According to the experimental results, the mAP value of the MFCMR method proposed in this paper is better than the baselines on dataset 1. The CCA method focuses on the association relationship between image text pairs, but it is difficult to establish a good association relationship for the same semantic data in different



(a) On dataset 1

(b) On dataset 2

Fig. 2. The performance of MFCMR with different method parameter k on different datasets.



(a) On dataset 1

(b) On dataset 2

Fig. 3. MFCMR's embedding loss and adversarial loss variation curve during the training process on different datasets.

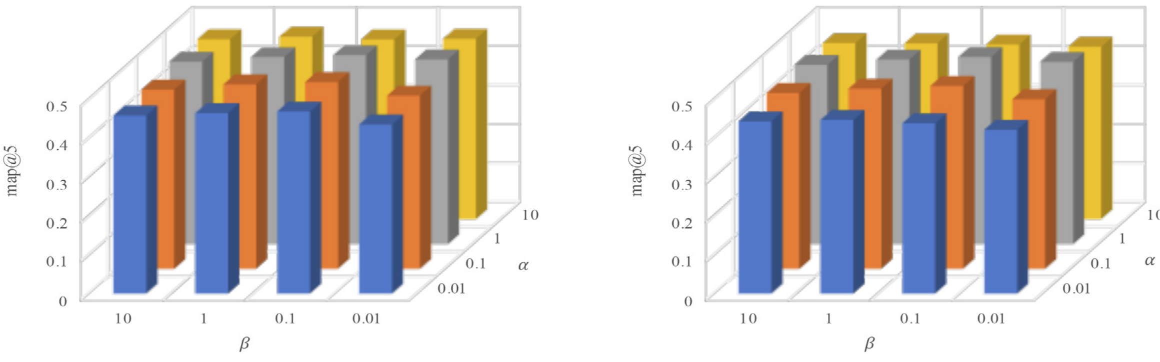
modalities. The JFSSL method increases the distance between different semantic data in the same modality, but it does not fully consider the distance between similar semantic data in different modalities, so it is difficult to describe the correlation between different modalities. The Corr-AE method simultaneously models the cross-media data association relationship and reconstruction information. The MFCMR method uses the adversarial learning method to make the semantic mapping network which can achieve a better structure, so the effect is better than the Corr-AE method. The JRL method uses graph-based conventions to describe semantic correlation and modality similarity. However, JRL is the traditional method, the effect is not as good as DNN-based methods. The Deep-SM method only considers coarse-grained global semantic information and does not consider the distribution differences between different modality data, so the effect is worse than MFCMR.

The ACMR method uses adversarial learning to make certain optimizations, however, it does not integrate the various features of scientific and technological information. So the effect of the ACMR method is not as good as MFCMR. The BATCH method is a supervised cross-modal hashing method. The experimental results show a trend of increasing mAP values as the amount of binary encoded data increases. However, converting cross-media semantic distances to binary still loses a lot of semantic information and therefore the effect is poor on dataset 1. The SSACR method uses real numbers to express the correlation between text data and image data, but it does not make full use of the multi-feature characteristics of scientific and technological information. This method, meanwhile, does not pay attention to semantic consistency regularization, so MFCMR can get better effects. The PAN method leverages a unified prototype to represent each semantic category across modalities and takes unified prototypes as anchors to learn cross-modal representations adaptively. However, the semantic regularization problem and the characteristics of scientific and technological information data are not taken into account, so the effect is still inferior to that of our method.

According to the analysis of the experimental results in Table 3, we can get results which is similar to the experimental results in Table 2. Firstly, because the sources of these data are the same, the data have similar data distributions. Secondly, although these data come from the same website, they are derived from scientific and technological information in different time periods. Therefore, the following conclusions can be drawn: MFCMR's cross-media retrieval effect on the scientific and technological information data set is better than other algorithms; On different data sets, the mAP value of this algorithm is better than baselines, which also proves the stability of the MFCMR.

One of the main parameters in the MFCMR method is the setting of the number of training times k of the feature mapping network in the adversarial training. Therefore, experiments are carried out in dataset 1 and dataset 2, and the mAP@5 value of the image search text and the text search image is used as the evaluation standard to judge the influence of the k value. The experimental results are shown in Fig. 2. The MFCMR method has slightly different values for the optimal k value on different datasets. When k is set to 4, the optimal effect is achieved. When k is greater than 4, the mAP value shows a slow downward trend. Therefore, in the case of determining the values of α and β, the value of $k$is 4 as the current optimal. Because the training times of the generator in adversarial training determine the effect of feature mapping to the common subspace, the value of $k$should not be too small; however, if the value of $k$is too large, the influence of the discriminator on the overall model training will be too neglected, thereby affecting the overall training effect.

In the MFCMR method, the adversarial learning method is used to optimize the embedding loss and adversarial loss of the objec-



(a) On dataset 1

(b) On dataset 2

Fig. 4. The average performance with different method parameters on different datasets.

Table 4
Performance of MFCMR method using different features to form L_{mod} on dataset 1.

	map@5			map@20			map@50		
	txt2img	img2txt	avg	txt2img	img2txt	avg	txt2img	img2txt	avg
MFCMR$_N$	0.5034	0.4592	0.4813	0.5074	0.4563	0.4819	0.4763	0.4338	0.4551
MFCMR$_T$	0.5054	0.4642	0.4848	0.5183	0.4578	0.4881	0.5068	0.4473	0.4771
MFCMR$_A$	0.5168	0.4683	0.4926	0.5253	0.4587	0.4920	0.5117	0.4434	0.4776
MFCMR	**0.5192**	**0.4703**	**0.4948**	**0.5307**	**0.4602**	**0.4955**	**0.5157**	**0.4520**	**0.4839**

Table 5
Performance of MFCMR method using different features to form L_{mod} on dataset 2

	map@5			map@20			map@50		
	txt2img	img2txt	avg	txt2img	img2txt	avg	txt2img	img2txt	avg
$\mathrm{MFCMR_N}$	0.4935	0.4587	0.4761	0.5039	0.4561	0.4800	0.4758	0.4364	0.4561
$\mathrm{MFCMR_T}$	0.4983	0.4593	0.4788	0.5064	0.4563	0.4814	0.4812	0.4394	0.4603
$\mathrm{MFCMR_A}$	0.5072	0.4662	0.4867	0.5216	0.4568	0.4892	0.4941	0.4407	0.4674
MFCMR	**0.5136**	**0.4670**	**0.4903**	**0.5221**	**0.4572**	**0.4897**	**0.4952**	**0.4501**	**0.4727**

tive function, and the effect of adversarial learning in MFCMR is further studied. The embedding loss and adversarial loss values of the first 60 epochs are recorded, as shown in the Fig. 3. We find that the loss function drops very fast, which shows that MFCMR is very easy to train.

When the value of k is fixed, we determine the influence of α and β on the overall search results in 16 sets of data. Experiments on dataset 1 and dataset 2 and the experimental results are shown in the Fig. 4. We find that MFCMR performs well when α and β are in the range of [0.1,1], and the mAP value will not fluctuate greatly with the changes of α and β.

MFCMR uses a multi-feature fusion method in feature mapping. To show the effectiveness of the time feature and author feature for training, this paper studies the impact on the overall result when these two features are not used ($\mathrm{MFCMR_N}$) and when a single feature is used ($\mathrm{MFCMR_T}$, $\mathrm{MFCMR_A}$). The experimental results are shown in Table 4 and Table 5.

As shown in Table 4, we can find that when the time and author characteristics are not used, the map values of dataset 1 are greatly reduced. When only the time feature is used, the mAP value of dataset 1 has a big drop; when only the author feature is used, the map value of dataset 1 drops slightly. Therefore, time feature and author feature have a relatively large impact on the overall search performance.

As shown in Table 5, we get a conclusion similar to Table 4, i.e., time feature and author feature are useful for cross-media retrieval of scientific and technological in-formation. We get more accurate results while integrating multiple features into cross-media retrieval. In addition, MFCMR, which combines time feature and author feature, also has obvious advantages on different datasets. The experiments on the two datasets both show that the MFCMR that combines multiple features has a higher mAP value, which proves that the MFCMR has better stability.

5. Conclusion

This paper proposes a Multi-feature Fusion based Cross-Media Retrieval (MFCMR). By learning the feature mapping network and the modality discrimination network, the spatial distance of data with the same semantics among different modalities can be reduced, and the spatial distance of data with different semantics can be increased. The feature mapping network maintains the semantic stability of the cross-modal data before and after the mapping, and the modality discrimination network reduces the modality difference of the cross-modal data so that the cross-media data is mapped into the same common subspace. This paper combines the unique characteristics of scientific and technological information and uses the MFCMR method based on multiple features. Finally, according to the trained model, the cross-media retrieval results are obtained. This paper uses multiple sets of experiments to verify that MFCMR achieves better performance for scientific and technological information data in different datasets.

CRediT authorship contribution statement

Yang Jiang: Methodology, Software, Writing - original draft. ***Junping Du:*** Writing - review & editing. ***Zhe Xue:*** Writing - review & editing. ***Ang Li:*** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by National Key R&D Program of China (2018YFB1402600), by the National Natural Science Foundation of China (61772083,61802028), and by Science and Technology Major Project of Guangxi (GuikeAA18118054).

References

[1] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering 26 (1) (2014) 97–107.
[2] Y. Yang, J. Du, Y. Ping, Ontology-based intelligent information retrieval system 26 (7) (2015) 1675–1687.
[3] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural computation 16 (12) (2004) 2639–2664.
[4] A. Li, J. Du, F. Kou, Z. Xue, X. Xu, Y. Jiang, Scientific and technological information oriented semantics-adversarial and media-adversarial cross-media retrieval (2022).
[5] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis (2013) 1247–1255.
[6] F. Yunman, M. Jianxia, Review on the lda-based techniques detection for the field emerging topic, Data Analysis and Knowledge Discovery 12 (2013) 58–65.
[7] B. Sun, J. Du, T. Gao, Study on the improvement of k-nearest-neighbor algorithm 4 (2009) 390–393.
[8] Z. Xue, J. Du, D. Du, S. Lyu, Deep low-rank subspace ensemble for multi-view clustering, Information Sciences 482 (2019) 210–227.
[9] P. Tar, N. Thacker, S. Deepaisarn, J. O'Connor, A. McMahon, A reformulation of plsa for uncertainty estimation and hypothesis testing in bio-imaging, Bioinformatics 36 (13) (2020) 4080–4087.
[10] R. Nallapati, W.W. Cohen, Link-plsa-lda: A new unsupervised model for topics and influence of blogs. (2008) 84–92.
[11] F. Kou, J. Du, Y. He, L. Ye, Social network search based on semantic analysis and learning, CAAI Transactions on Intelligence Technology 1 (4) (2016) 293–302.
[12] C. Yang, B. Zhang, R. Li, Guo.Qiang, Topic discovery and clustering for online journals based on lda algorithm, University of Shanghai for Science and Technology 41 (3) (2019) 273–280.
[13] M. Li, J. Liu, Q. Guo, R. Li, X. Tang, Topic discovery and clustering research for online courses based on text mining, University of Shanghai for Science and Technology 40 (3) (2018) 259–266.
[14] X. Pang, B. Wan, P. Wang, Topic mining for microblog based on mb-lda model, Journal of Computer Research and Development 44 (8) (2017) 236–241.
[15] W. Hu, J. Gao, B. Li, O. Wu, J. Du, S. Maybank, Anomaly detection using local kernel density estimation and context-based regression, IEEE Transactions on Knowledge and Data Engineering 32 (2) (2018) 218–233.
[16] Q. Li, J. Du, F. Song, C. Wang, H. Liu, C. Lu, Region-based multi-focus image fusion using the local spatial frequency (2013) 3792–3796.
[17] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arxiv, arXiv preprint arXiv:1409.1556.

[18] P. Dhankhar, Resnet-50 and vgg-16 for recognizing facial emotions, International Journal of Innovations in Engineering and Technology (IJIET) 13 (4) (2019) 126–130.
[19] H. Qassim, A. Verma, D. Feinzimer, Compressed residual-vgg16 cnn model for big data places image recognition (2018) 169–175.
[20] H. Yuan, J. Li, L.L. Lai, Y.Y. Tang, Low-rank matrix regression for image feature extraction and feature selection, Information Sciences 522 (2020) 214–226.
[21] X. Peng, X. Zhang, Y. Li, B. Liu, Research on image feature extraction and retrieval algorithms based on convolutional neural network, Journal of Visual Communication and Image Representation 69 (2020) 102705.
[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets 27 (2014).
[23] Y. Fang, W. Deng, J. Du, J. Hu, Identity-aware cyclegan for face photo-sketch synthesis and recognition, Pattern Recognition 102 (2020) 107249.
[24] C. Shi, X. Han, L. Song, X. Wang, S. Wang, J. Du, P.S. Yu, Deep collaborative filtering with multi-aspect information in heterogeneous networks, IEEE Transactions on Knowledge and Data Engineering 33 (4) (2021) 1413–1425.
[25] W. Li, Y. Jia, J. Du, Recursive state estimation for complex networks with random coupling strength, Neurocomputing 219 (2017) 1–8.
[26] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval (2017) 154–162.
[27] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, H.T. Shen, Unsupervised cross-modal retrieval through adversarial learning (2017) 1153–1158.
[28] Y. Peng, J. Qi, Cm-gans: Cross-modal generative adversarial networks for common representation learning, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15 (1) (2019) 1–24.
[29] S. Chun, S.J. Oh, R.S. De Rezende, Y. Kalantidis, D. Larlus, Probabilistic embeddings for cross-modal retrieval (2021) 8415–8424.
[30] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (4) (2021) 1–23.
[31] Y. Wang, X. Luo, L. Nie, J. Song, W. Zhang, X.-S. Xu, Batch: A scalable asymmetric discrete cross-modal hashing, IEEE Transactions on Knowledge and Data Engineering 33 (11) (2020) 3507–3519.
[32] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10) (2015) 2010–2023.
[33] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder (2014) 7–16.
[34] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, IEEE Transactions on Circuits and Systems for Video Technology 24 (6) (2013) 965–978.
[35] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with cnn visual features: A new baseline, IEEE Transactions on Cybernetics 47 (2) (2016) 449–460.
[36] C. Liu, J. Du, N. Zhou, A cross media search method for social networks based on adversarial learning and semantic similarity, Science China Information Sciences 51 (5) (2021) 779–794.
[37] Z. Zeng, S. Wang, N. Xu, W. Mao, Pan: Prototype-based adaptive network for robust cross-modal retrieval (2021) 1125–1134.

Yang Jiang was born in 1995, is a Master candidate in Computer Science of Beijing University of Posts and Telecommunications. His research interests include nature language processing, cross-modal retrieval and deep learning.

Junping Du was born in 1963. She is now a professor and Ph.D. tutor at the School of Computer Science and Technology, Beijing University of Posts and Telecommunications. Her research interests include artificial intelligence, machine learning and pattern recognition.

Zhe Xue received the Ph.D. degree in computer science from University of Chinese Academy of Sciences, Beijing, China in 2017. He is currently an associate professor with the school of computer science, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include machine learning, data mining and multimedia data analysis.

Ang Li received the B.S. degree from the Nanchang Hangkong University, China, in 2015 and the M.S. degree from the Beijing University of Posts and Telecommunications, China, in 2019, all related to computer science. He is currently working toward the Ph.D. degree in Computer Science and Technology at the Beijing University of Posts and Telecommunications, China. His major research interests include information retrieval and data mining.