



Representation separation adversarial networks for cross-modal retrieval

Jiaxin Deng¹ · Weihua Ou^{1,2,4} · Jianping Gou³ · Heping Song³ · Anzhi Wang¹ · Xing Xu⁵

Published online: 5 June 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Cross-modal retrieval aims to search the semantically similar instances from the other modalities by giving a query from one modality. Recently, generative adversarial networks (GANs) has been proposed to model the joint distribution over the data from different modalities and to learn the common representations for cross-modal retrieval. However, most of existing GANs-based methods simply project original representations of different modalities into a common representation space, and ignore the fact that different modalities share the common characteristics and on the other side each modality has the individual characteristics. To address this problem, in this paper, we propose a novel cross-modal retrieval method, called representation separation adversarial networks, which explicitly separates the original representations into common latent representations and private representations. Specifically, we minimize the correlation between the common representations and private representations to ensure independence of them. Then, we reconstruct the original representations via exchanging the common representations of different modalities to encourage the information swap. Finally, the labels are utilized to increase the discriminant of common representations. Comprehensive experimental results on two widely used datasets show that the proposed method achieved better performance than many existing GANs-based methods, and demonstrate that explicitly modeling the private representation for each modality can improve the model to extract common latent representations.

Keywords Cross-modal retrieval · Adversarial learning · Common representation · Private representation · Representation separation

1 Introduction

In social network, lots of multi-modal data, such as, image, video, text, audio are mixed together and endow semantic correlations. Many single-modality approaches have been proposed to understand those data, such as image classification or retrieval [1, 2], sentence semantic matching [3] and answer selection [4, 5]. There is immediate need to analyze those data across different modalities, such as retrieving similar instances from the other modalities giving a query from one modality [6], i.e., cross-modal retrieval. In recent years, cross-modal retrieval has gaining lots of attentions [7–10]. The main challenge for cross-modal retrieval is how to measure the similarity between the data from different modalities because of the semantic gap, heterogeneity and diversity within them.

To mitigate this problem, an intuitive way is to learn a common latent representation space, in which the

✉ Weihua Ou
ouweihuahust@gmail.com

¹ School of Big Data and Computer Science, Guizhou Normal University, Guiyang, People's Republic of China

² Special Key Laboratory of Artificial Intelligence and Intelligent Control of Guizhou Province, Guiyang, People's Republic of China

³ School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, Jiangsu, People's Republic of China

⁴ Research Center of Security Video and Image Processing Engineering Technology of Guizhou, Guiyang, People's Republic of China

⁵ School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, People's Republic of China

similarity between data from different modalities can be measured directly. For example, the classical methods are to learn linear projection by maximizing the correlations between the pair-wise data from different modalities. Such as, canonical correlation analysis (CCA) [11] was adopted to project text features and image features into a low-dimensional common subspace for cross-modal retrieval [12]. Furthermore, some extensions of kernel-based methods [13, 14] have been proposed to model more complex correlations among different modalities. However, the main drawback of those methods is that they are simply to project the original representation into a common space and neglect the unique property of each modality.

Thanks to the successful applications of deep neural network (DNN), a large number of DNN-based methods have been proposed to cross-modal retrieval. For instance, deep canonical correlation analysis (DCCA) [15] combined deep network with CCA for cross-modal retrieval. The correspondence autoencoder (Corr-AE) [16] was proposed to model the correlations of different modalities through incorporating representation learning and correlation learning. Cross-modal multiple deep networks (CMDN) [17] was proposed by constructing hierarchical network structure to model the correlations of inter-modality and intra-modality.

Recently, some GANs-based cross-modal retrieval methods [6, 18] have been proposed. For example, adversarial cross-modal retrieval (ACMR) [6] proposes to learn the discriminative and modality-invariant common representations by adversarial learning. Cross-modal generative adversarial networks for common representation learning (CM-GANs) [18] exploits the cross-modal correlation by the weight-sharing constraint. However, most of the existing DNN-based [16, 17, 19–22] and GANs-based works [6, 18] simply project original representations into a common representation space and ignore the specific information in each modality. In fact, data from different modalities have some common characteristics and also have private characteristics for each modality. Because those data come from different modalities and have inconsistent distributions. An intuitive way is to introduce a private subspace to capture modality specific properties, and introduce a common subspace to capture the properties shared by different modalities [23].

In this paper, we propose a novel cross-modal retrieval method, called Representation Separation Adversarial Networks (ReSAN), which separates the original representations into common representation and private representation. Figure 1 shows the framework, which includes two sub-networks, i.e., image sub-networks and text sub-networks. First, to separate the original representation, we minimize the correlations between common and private representation to encourage them to be independent. At

shown in the red box with dotted lines in the Fig. 1, we hope that common representations only contains the components shared by different modalities, while private representations only contains the unique components of each modality. Then, we reconstruct the original representations by exchanging the common representation to encourage the information swap across modalities. Finally, we use semantic information to make the common representation to be discriminative and modality-invariant. The main contributions of this work can be summarized as follows:

- We propose a representation separation adversarial networks for cross-modal retrieval, which explicitly splits the original representations into common representations and private representations.
- We propose a modality-variant common representation learning strategy, which can exchange the information among modalities during learning processing.
- We evaluate the proposed method ReSAN on cross-modal retrieval and the results demonstrate it obtained best performance compared to most existing methods.

The rest of paper is organized as follows. We first briefly review the related works in Sect. 2, and present the proposed method in Sect. 3. Then, we derive the algorithm in Sect. 4, and conduct experiments in Sect. 5. Finally, we conclude this paper in Sect. 6.

2 Related works

Since generative adversarial networks (GANs) [24] have been proposed in 2014, it has been used in a wide applications, such as image style transformation [25, 26], image synthesis [27, 28], object tracking [29] and zero-shot learning [30, 31]. The original GANs consists of a generative model G and a discriminative model D . The generative model aims to generate fake data and capture the distribution over real data, discriminative model aims to discriminate the real data and generated data. G and D play the minimax game on $V(G, D)$ as follows:

$$\min_G \max_D V(G, D) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

where x denotes the real data and z is the noise input. Wang et al. [6] first introduce the GANs into cross-modal retrieval and proposed adversarial cross-modal retrieval (ACMR), which learns modality-invariant and discriminative common representations through adversarial learning. Wu et al. [32] proposed cycle-consistent deep generative hashing for cross-modal retrieval, which learns a couple of hash mappings by cycle-consistent adversarial learning without paired input-output examples. Peng

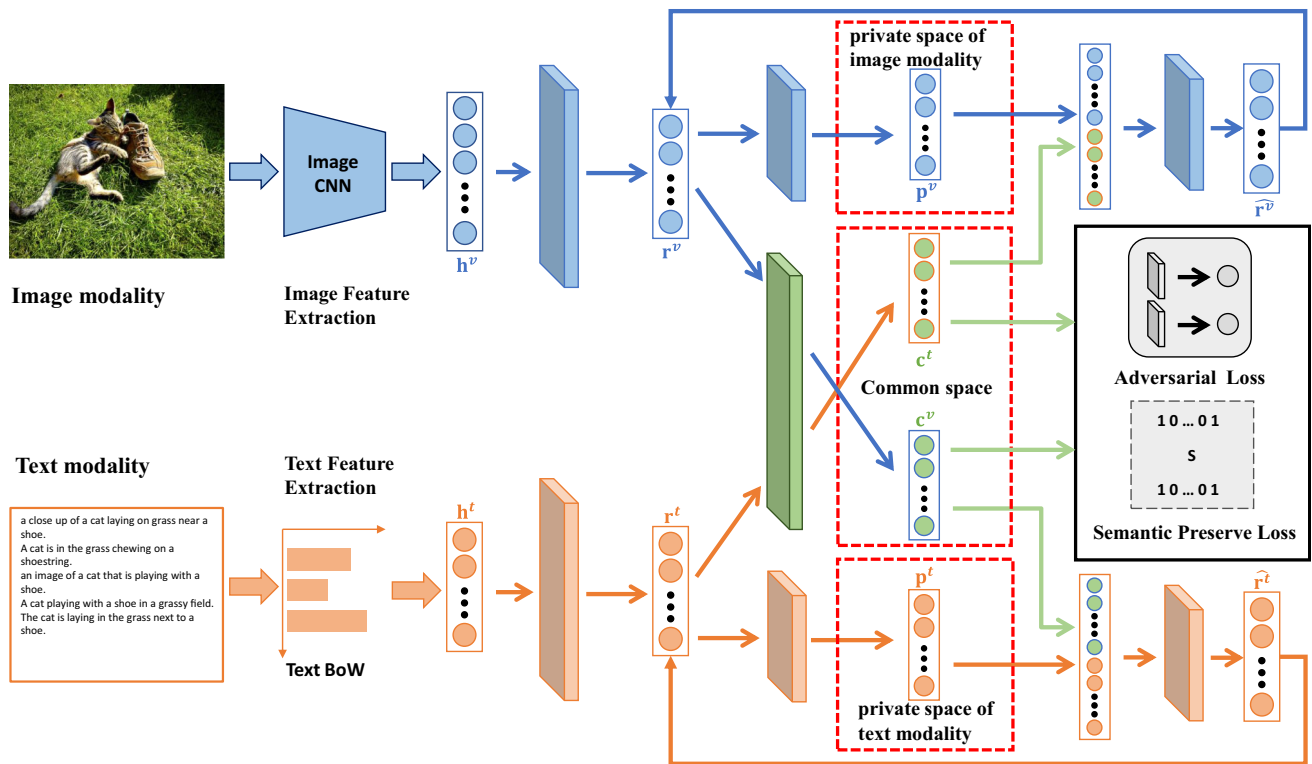


Fig. 1 The flowchart of the proposed representation separation adversarial networks (ReSAN) for cross-modality retrieval, which includes two sub-networks. The upper sub-network is the image

representation learning network, while the below one is the text representation learning network

et al. [18] proposed cross-modal generative adversarial networks for common representation learning (CM-GANs), which considers both inter-modality and intra-modality adversarial learning in a more effective manner to learn common representation. Although those methods exploit the cross-modal correlation by adversarial, however, they ignore private components of the original representations.

Recently, domain separation networks (DSNs) [33] is proposed for transfer learning [34]. It explicitly separate the representations of different domain into two parts: one is the private component and the other is the shared component across domains. The experimental results demonstrate its success in unsupervised domain adaption scenario. Yang et al. [23] propose shared predictive cross-modal deep quantization (SPDQ), which construct a shared subspace and two private subspaces to adequately exploit the intrinsic correlations among multiple modalities. Inspired by those works, this paper is dedicated to separate the original representations into common representation and private representation explicitly and exploit common latent semantic representations. Different from works [23], we achieve this idea under the framework of generative adversarial networks.

3 Proposed method

3.1 Notations and problem statement

3.1.1 Notations

To simplify the notations, we focus on two modalities, i.e., image modality and text modality. Assuming N instances of image-text pairs, we denote the whole dataset as $\mathcal{O} = \{(\mathbf{v}_i, \mathbf{t}_i)\}_{i=1}^N$, where \mathbf{v}_i is the i -th image feature vector, \mathbf{t}_i is the i -th text feature vector. For each pair of data $(\mathbf{v}_i, \mathbf{t}_i)$, the semantic label is assigned by vector $\mathbf{y}_i = [y_i^1, y_i^2, \dots, y_i^d]$, where d is the total number of categories, $y_i^j = 1$ if $(\mathbf{v}_i, \mathbf{t}_i)$ belongs to the j -th class while $y_i^j = 0$ otherwise.

3.1.2 Problem statement

Since the image feature vectors and text feature vectors typically have different statistical properties, they cannot be directly compared. To address this problem, we propose to learn the transform function $\mathbf{c}_i^v = f(\mathbf{v}_i; \mathcal{Y}_v)$ for image modality, and $\mathbf{c}_i^t = g(\mathbf{t}_i; \mathcal{Y}_t)$ for text modality, respectively. \mathbf{c}_i^v and \mathbf{c}_i^t denote image and text representation in common representation space, \mathcal{Y}_v and \mathcal{Y}_t are the parameters of the

two functions, respectively. After that, we can measure their similarity by calculating the cosine distance between \mathbf{c}_i^v and \mathbf{c}_i^t . We expect the cosine distance of the semantically similar image-text pairs is smaller than that of the semantically dissimilar image-text pairs.

3.2 Model

Inspired by GANs’ strong ability in modelling data distribution and learning discriminative representations, we utilize GANs to model the distribution over the data of different modalities and to learn the common representations. In this paper, we introduce two generative adversarial networks: GAN_v for image modality and GAN_t for text modality.

3.2.1 Generative model

Image representation Generator G_v and text representation Generator G_t take image-text paired features \mathbf{h}^v and \mathbf{h}^t as the inputs, respectively. Through several fully-connected layers, the same length representations \mathbf{r}^v and \mathbf{r}^t are obtained for image modality and text modality. Then, image representations \mathbf{r}^v are separated into common representation \mathbf{c}^v and private representation \mathbf{p}^v , while text representations \mathbf{r}^t are also separated into common representation \mathbf{c}^t and private representation \mathbf{p}^t , as shown in Fig. 1.

Ideally, we expect that the representations in common space only include the semantic information shared by images and texts, while the private representation only contains their own specific information. To achieve this, we minimize the cosine distances between representations in common space and private space for each modality. This can be formulated as follows:

$$L_{Space_*} = \frac{1}{K} \sum_{i=1}^K \langle \mathbf{c}_i^*, \mathbf{p}_i^* \rangle > 2 \tag{2}$$

where $* \in \{v, t\}$, represents different modalities, $\langle \mathbf{a}, \mathbf{b} \rangle$ is the inner product of \mathbf{a} and \mathbf{b} , K is the number of instance in one batch.

To ensure the effectiveness of separation and improve the information swap among modalities, we reconstruct the original representations by exchanging the common representations among modalities. Specifically, we concatenate image private representation \mathbf{p}^v and text common representations \mathbf{c}^t as the input of several fully-connected layers to reconstruct the image representations $\hat{\mathbf{r}}^v$. Similarly, we concatenate text private representation \mathbf{p}^t and image common representations \mathbf{c}^v as the input of several fully-connected layers to reconstruct text representations $\hat{\mathbf{r}}^t$. The reconstruction loss can be formulated as follows:

$$L_{Recon_*} = \frac{1}{K} \sum_{i=1}^K (\hat{\mathbf{r}}_i^* - \mathbf{r}_i^*)^2 \tag{3}$$

where $* \in \{v, t\}$.

To exploit the common semantics from inter-modality and intra-modality, we denote the inter-modality similarity matrix of image and text as S^{vv} and S^{tt} , the intra-modality similarity matrix as S^{vt} . For image modality S^{vv} , we set $S_{ij}^{vv} = 1$ if image v_i and image v_j are the same class, and $S_{ij}^{vv} = 0$ otherwise. Similarly, for text modality S^{tt} , $S_{ij}^{tt} = 1$ if text t_i and text t_j are the same class, and $S_{ij}^{tt} = 0$ otherwise. For S^{vt} , we define $S_{ij}^{vt} = 1$ if image v_i and text t_j belong to the same class, and otherwise $S_{ij}^{vt} = 0$.

Based on the above notations and discussion, the objective function of image modality is defined as follows:

$$L_{S_v} = -\frac{1}{K} \sum_{i,j=1}^K (S_{ij}^{vt} \Theta_{ij} - \log(1 + e^{\Theta_{ij}})) - \frac{1}{K} \sum_{i,j=1}^K (S_{ij}^{vv} \Gamma_{ij} - \log(1 + e^{\Gamma_{ij}})) \tag{4}$$

where $\Theta_{ij} = \cos(\mathbf{c}_i^v, \mathbf{c}_j^t)$, $\Gamma_{ij} = \cos(\mathbf{c}_i^v, \mathbf{c}_j^v)$, $\cos(\mathbf{a}, \mathbf{b})$ is the cosine function used to compute the similarity between \mathbf{a} and \mathbf{b} .

Similarly, the objective function of text modality can be formulated as follows:

$$L_{S_t} = -\frac{1}{K} \sum_{i,j=1}^K (S_{ij}^{vt} \Theta_{ij} - \log(1 + e^{\Theta_{ij}})) - \frac{1}{K} \sum_{i,j=1}^K (S_{ij}^{tt} \Phi_{ij} - \log(1 + e^{\Phi_{ij}})) \tag{5}$$

where $\Phi_{ij} = \cos(\mathbf{c}_i^t, \mathbf{c}_j^t)$. The first term in equations (4) and (5) is the negative log likelihood of the cross-modal similarities with the likelihood function defined as follows:

$$P(S_{ij}^{vt} | \mathbf{c}_i^v, \mathbf{c}_j^t) = \begin{cases} \sigma(\Theta_{ij}) & \text{when } S_{ij}^{vt} = 1 \\ 1 - \sigma(\Theta_{ij}) & \text{when } S_{ij}^{vt} = 0 \end{cases}$$

where $\sigma(\Theta_{ij}) = \frac{1}{1 + e^{-\Theta_{ij}}}$.

It is easy to find that minimizing this negative log likelihood is equivalent to maximize the likelihood, which can make the similarity between \mathbf{c}_i^v and \mathbf{c}_j^t to be large when $S_{ij}^{vt} = 1$, and to be small when $S_{ij}^{vt} = 0$. The second term in (4) and (5) measure the inter-similarity of the image and text modality, respectively. Therefore, Eqs. (4) and (5) encourage to learn more discriminative common representations.

3.2.2 Discriminative model

Two discriminators are designed to distinguish the representations from common representation space. The image representation discriminator D_v tries to distinguish the image representations \mathbf{c}^v as the real data from representations \mathbf{c}^t as fake data. The text discriminator D_t tries to distinguish the representations \mathbf{c}^t as the real data from the representations \mathbf{c}^v as fake data. Based on this, the adversarial loss for image modality can be defined as follow:

$$L_{adv_v} = E_{\mathbf{c}^v \sim P_{c^v}} [\log(D_v(\mathbf{c}^v))] + E_{\mathbf{c}^t \sim P_{c^t}} [\log(1 - D_v(\mathbf{c}^t))] \quad (6)$$

Similarly, adversarial loss for text modality can be defined as follow:

$$L_{adv_t} = E_{\mathbf{c}^t \sim P_{c^t}} [\log(D_t(\mathbf{c}^t))] + E_{\mathbf{c}^v \sim P_{c^v}} [\log(1 - D_t(\mathbf{c}^v))] \quad (7)$$

After the adversarial learning, ultimately the discriminators cannot identify the representations come from which modality. The cross-modal correlations could be well learned and the discriminative common properties are simultaneously captured.

3.3 Objective function

With the above definitions, the whole objective function can be formulated as follows:

$$\min_{G^v, G^t, D_v, D_t} \max_{G^v, G^t, D^v, D^t} L_{GAN_v}(G^v, G^t, D^v) + L_{GAN_t}(G^v, G^t, D^t) \quad (8)$$

where $L_{GAN_*} = L_{adv_*} + \alpha L_{Space_*} + \beta L_{Recon_*} + \gamma L_{S_*}$, $*$ $\in \{v, t\}$, α , β and γ are the regularization parameters.

4 Algorithm

4.1 Optimizing discriminative model

Following [6], we adopt stochastic gradient method to optimize the discriminator. For the image pathway, image discriminator is conducted to maximize the log-likelihood to correctly discriminate common representations. It is trained by ascending their stochastic gradient with the following equation:

$$\theta_{D_v} \leftarrow \theta_{D_v} + \mu \cdot \nabla_{\theta_{D_v}} \frac{1}{K} \sum_{i=1}^K [\log(D_v(\mathbf{c}_i^v)) + \log(1 - D_v(\mathbf{c}_i^t))] \quad (9)$$

where θ_{D_v} are parameters of image discriminative model, μ is learning rate. Similarly, for the text discriminator in text pathway, it is trained by ascending their stochastic gradient with the following equation:

$$\theta_{D_t} \leftarrow \theta_{D_t} + \mu \cdot \nabla_{\theta_{D_t}} \frac{1}{K} \sum_{i=1}^K [\log(D_t(\mathbf{c}_i^t)) + \log(1 - D_t(\mathbf{c}_i^v))] \quad (10)$$

where θ_{D_t} are parameters of text discriminative model.

4.2 Optimizing generative model

For the image generator, it is trained by descending their stochastic gradient with the following equation:

$$\theta_{G_v} \leftarrow \theta_{G_v} - \mu \cdot \nabla_{\theta_{G_v}} \left[\frac{1}{K} \sum_{i=1}^K \log(D_t(\mathbf{c}_i^v)) + \alpha L_{Space_v} + \beta L_{Recon_v} + \gamma L_{S_v} \right] \quad (11)$$

For the text generator, similarly, it is updated parameters by descending the stochastic gradient as follows:

$$\theta_{G_t} \leftarrow \theta_{G_t} - \mu \cdot \nabla_{\theta_{G_t}} \left[\frac{1}{K} \sum_{i=1}^K \log(D_v(\mathbf{c}_i^t)) + \alpha L_{Space_t} + \beta L_{Recon_t} + \gamma L_{S_t} \right] \quad (12)$$

The details of the whole procedure is summarised in Algorithm 1.

Algorithm 1 Pseudocode of the proposed ReSAN

Require: The training data set $\mathcal{O}_{tr} = \{(\mathbf{v}_i, \mathbf{t}_i)\}_{i=1}^{N_{tr}}$, and label matrix $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \in \mathbb{R}^{d \times N_{tr}}$, parameters α, β and γ , the learning rate μ , minibatch size K for each iteration.

Repeat:

Sample matching image and text pairs $\{\mathbf{v}_i\}_{i=1}^K$ and $\{\mathbf{t}_i\}_{i=1}^K$;

Image pathway generation $G_v(\mathbf{v}_i) \rightarrow (\mathbf{r}_i^v, \mathbf{c}_i^v, \mathbf{p}_i^v)$;

Text pathway generation $G_t(\mathbf{t}_i) \rightarrow (\mathbf{r}_i^t, \mathbf{c}_i^t, \mathbf{p}_i^t)$;

Update parameters of image discriminative model θ_{D_v} by equation: $\theta_{D_v} \leftarrow \theta_{D_v} + \mu \cdot \nabla_{\theta_{D_v}} \frac{1}{K} \sum_{i=1}^K [\log(D_v(\mathbf{c}_i^v)) + \log(1 - D_v(\mathbf{c}_i^t))]$;

Update parameters of text discriminative model θ_{D_t} by equation: $\theta_{D_t} \leftarrow \theta_{D_t} + \mu \cdot \nabla_{\theta_{D_t}} \frac{1}{K} \sum_{i=1}^K [\log(D_t(\mathbf{c}_i^t)) + \log(1 - D_t(\mathbf{c}_i^v))]$;

for s steps **do**

Sample matching image and text pairs;

Update parameters of image generative model θ_{G_v} by equation: $\theta_{G_v} \leftarrow \theta_{G_v} - \mu \cdot \nabla_{\theta_{G_v}} \left[\frac{1}{K} \sum_{i=1}^K \log(D_t(\mathbf{c}_i^v)) + \alpha L_{Space_v} + \beta L_{Recon_v} + \gamma L_{S_v} \right]$;

Update parameters of text generative model θ_{G_t} by equation: $\theta_{G_t} \leftarrow \theta_{G_t} - \mu \cdot \nabla_{\theta_{G_t}} \left[\frac{1}{K} \sum_{i=1}^K \log(D_v(\mathbf{c}_i^t)) + \alpha L_{Space_t} + \beta L_{Recon_t} + \gamma L_{S_t} \right]$;

end for

Until: GANs converges;

5 Experiments and results

To evaluate the proposed method, we conduct experiments on the Wikipedia and the NUSWIDE-10k datasets. In Sect. 5.1, we describe the datasets and evaluation, following by the implementations details in Sect. 5.2. In Sect. 5.3, we show the experimental results and analysis.

5.1 Datasets and evaluation

5.1.1 Datasets

The Wikipedia [35] and the NUSWIDE-10k [36] datasets are widely used for cross-modal retrieval. The Wikipedia dataset consists of 10 categories, 2866 instances (image-text pairs), in which 2173 image-text pairs are randomly selected for training, the rest of 693 image-text pairs are for testing. The NUSWIDE-10k includes 10 categories and contains 10,000 image-text pairs, in which 8000 image-text pairs are randomly selected for training and the 2000 image-text pairs are selected for testing. We adopt 4096d vector extracted by the fc7 layer of VGGNet [37] as image feature, text features are 3000d bag-of-words (BoW) vector in Wikipedia and 1000 BoW in NUSWIDE-10k dataset.

The statistical results of those two datasets are summarised in Table 1.

5.1.2 Evaluation

In this paper, we use mean Average Precision (mAP) to evaluate the cross-modal retrieval performances.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP(q_i)$$

where $AP(\cdot)$ computes the average precision, N is the number of query samples and q_i represents the i -th query sample. The larger the mAP value is, the better the retrieval performance is. We conduct two different tasks including retrieving text using image as query (Img2Txt) and retrieving image using text as query (Txt2Img), and report the performance of mAP. The results of ACMR are obtained by implementing the code provided by the authors, and the others are reported from the published papers.

5.1.3 Compared methods

To show the effectiveness of our method, we selected following representative methods for comparison,

Table 1 The details of two datasets, where “/” in column “Instances” represents the number of training/test image-text pairs

Dataset	Instances	Categories	Image feature	Text feature
Wikipedia	2173/462	10	4096d VGG	3000d BoW
NUSWIDE-10k	8000/1000	10	4096d VGG	1000d BoW

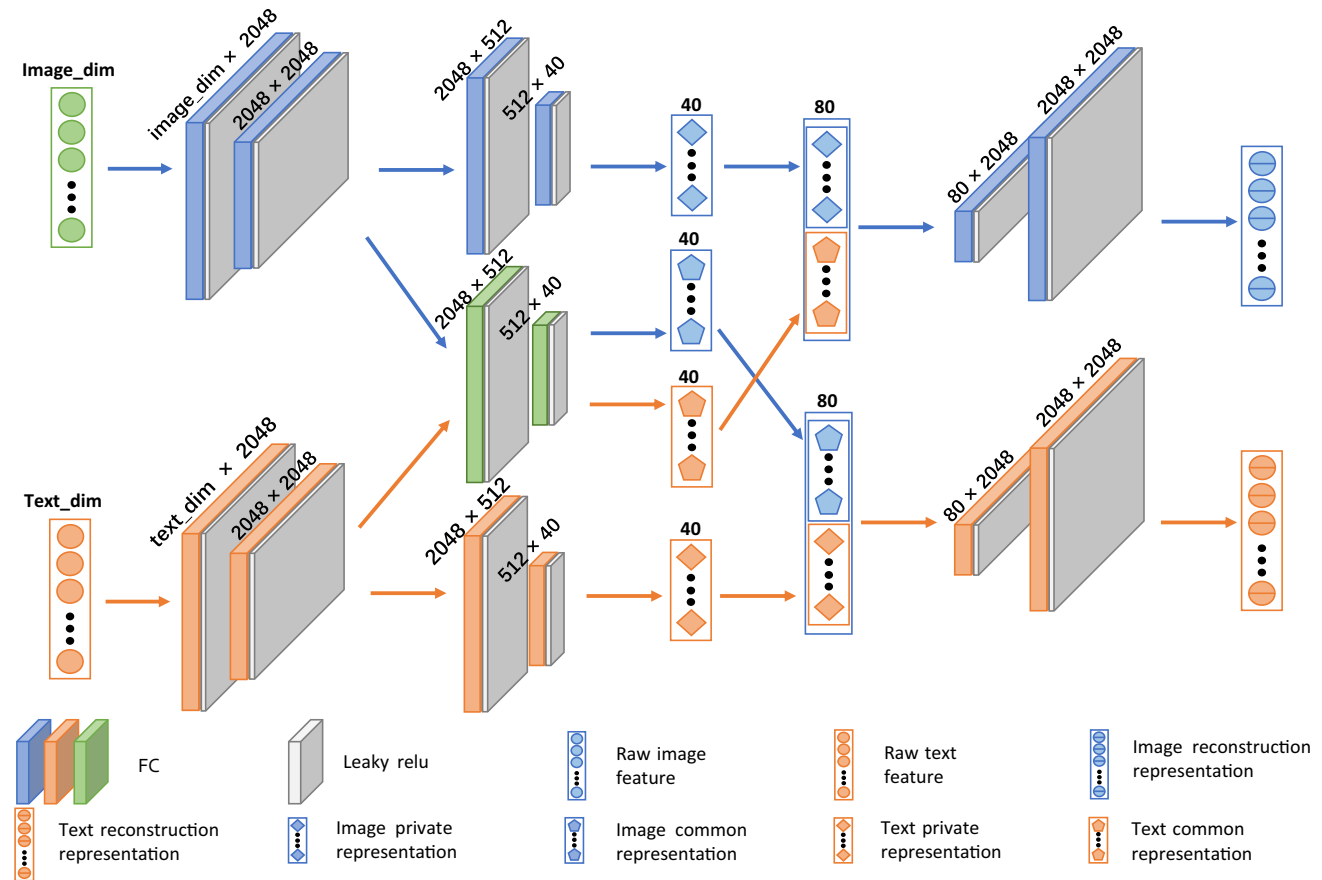


Fig. 2 The structural details of the representation separation adversarial networks (ReSAN)

Table 2 The mAP of different methods on the Wikipedia dataset for Img2Txt and Txt2Img

Methods	Img2Txt	Txt2Img	Avg.
CCA [38]	0.267	0.222	0.245
LCFS [39]	0.455	0.398	0.427
JRL [40]	0.453	0.400	0.426
Bimodal-AE [41]	0.314	0.290	0.302
Corr-AE [16]	0.402	0.395	0.398
CMDN [17]	0.488	0.427	0.458
ACMR [6]	0.515	0.471	0.493
ReSAN (ours)	0.526	0.487	0.507

The best accuracy values are given in bold

including traditional methods, DNN-based methods and GANs-based methods.

Traditional methods:

- CCA [38]: It learns linear projection by maximizing the correlation between pairwise data of different modalities.
- LCFS [39]: It learns two projection matrices with sparsity penalties to select relevant and discriminative

features from the coupled feature spaces simultaneously.

- JRL [40]: It integrates graphs regularization and semi-supervised information to jointly learn representation for different modalities.

DNN-based methods:

- Bimodal-AE [41]: It proposes a novel application of deep networks to learn features over multiple modalities.
- Corr-AE [16]: It models jointly the cross-modal correlation and reconstruction information.
- CMDN [17]: It models inter-modal invariance and intra-modal discrimination jointly in a multi-task learning framework.

GANs-based method:

- ACMR [6]: It seeks an effective common subspace based on adversarial learning.

Table 3 The mAP of different methods on the NUSWIDE-10k dataset for Img2Txt and Txt2Img

Methods	Img2Txt	Txt2Img	Avg.
CCA [38]	0.189	0.188	0.189
LCFS [39]	0.383	0.346	0.365
JRL [40]	0.426	0.376	0.401
Bimodal-AE [41]	0.327	0.369	0.348
Corr-AE [16]	0.366	0.417	0.392
CMDN [17]	0.492	0.515	0.504
ACMR [6]	0.529	0.547	0.538
ReSAN (Ours)	0.536	0.569	0.553

The best accuracy values are given in bold

5.2 Implementation details

The proposed method consists of two sub-networks, one for image modality and the other for text modality. As shown in the Fig. 2, the generative model of image modality is a four fully connected layer network with Leaky ReLU activation function, which projects the raw image features into a common subspace. We use a fully connected layer map the 2048 dimensional vector from the middle layer to the image private representation space. Then, we concatenate image private representation and text common representation to reconstruct the image representation with a fully connected layer. The generative model for text modality is similar to image modality. Each discriminative

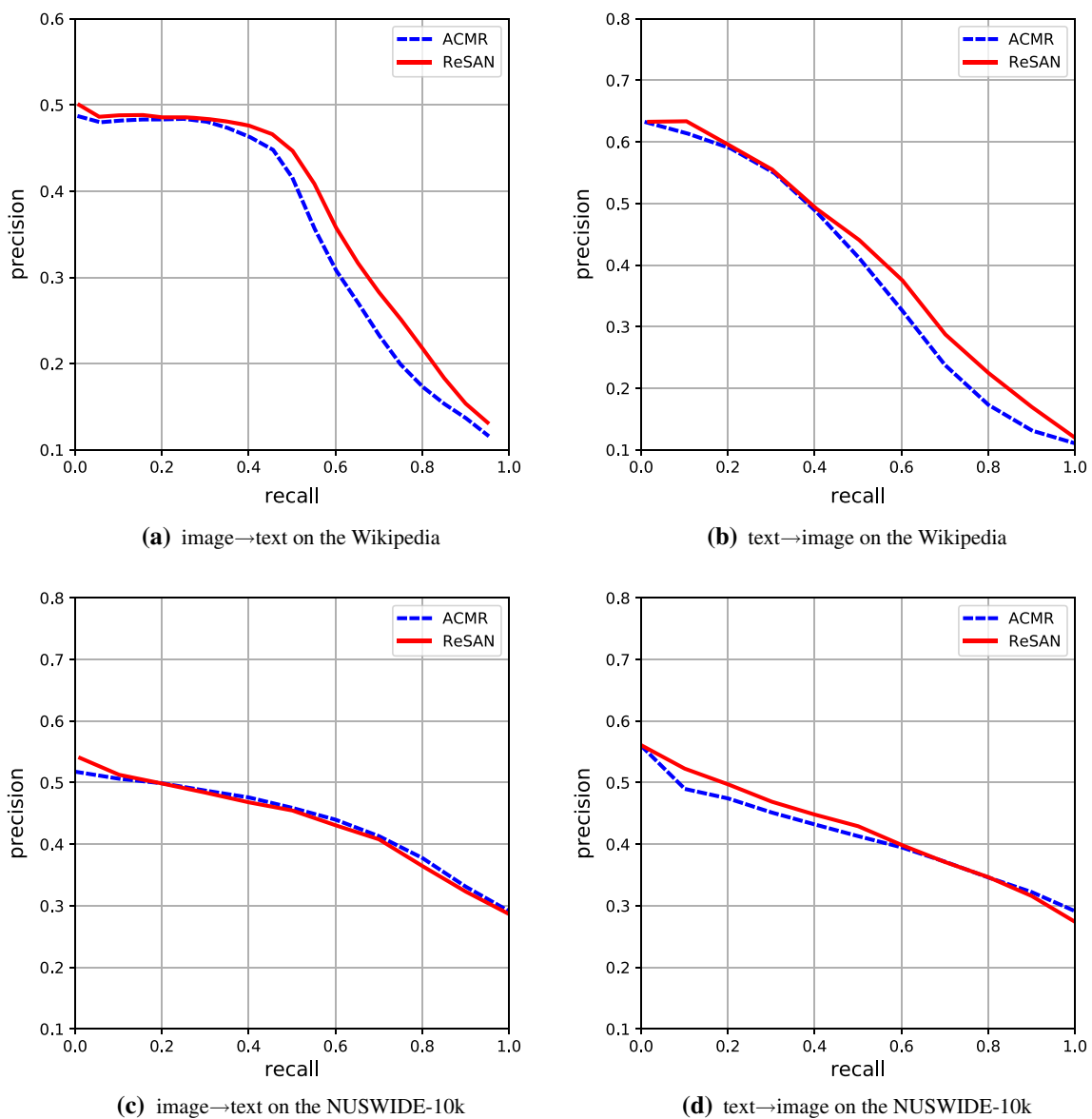


Fig. 5 The PR curves of ACMR and ReSAN on Wikipedia and NUS-WIDE-10k dataset

Table 4 The contribution for different terms in ReSAN on the Wikipedia and NUSWIDE-10k dataset

Method	Wikipedia			NUSWIDE-10k		
	Img2txt	Txt2img	Avg.	Img2txt	Txt2img	Avg.
ReSAN-P	0.512	0.462	0.487	0.471	0.505	0.488
ReSAN-D	0.475	0.437	0.456	0.517	0.542	0.530
ReSAN-C	0.511	0.472	0.492	0.529	0.557	0.543
ReSAN	0.526	0.487	0.507	0.536	0.569	0.553

The best accuracy values are given in bold

model consists of two fully connected layers: the neurons number in the first layer is 128, and the second layer is 1. The mini-batch size is set to 64. Moreover, the α , β and γ are empirically set to 0.1, 0.1 and 1, respectively.

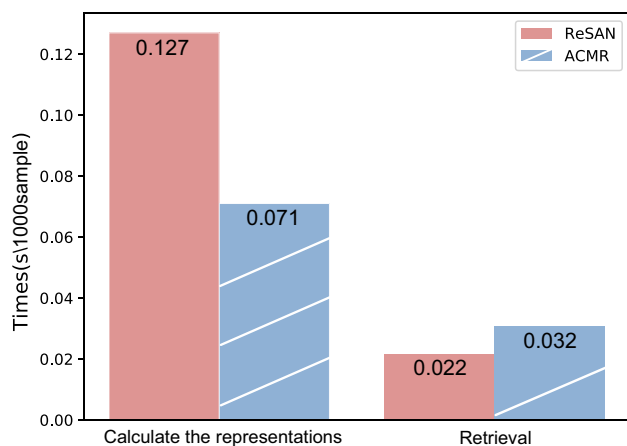
5.3 Experimental results

5.3.1 Results on Wikipedia dataset

Table 2 shows the mAP of different methods on this dataset. From that, we can see that LCFS and JRL obtained the similar results, which are much better than the traditional CCA method. GAN-based methods such as ACMR and ReSAN obtained better results than other traditional methods and DNN-based methods. Among them, our method achieved the best performance on this dataset.

Furthermore, we show the top-10 results retrieved by CCA, ACMR, and ReSAN on Wikipedia dataset in Fig. 3.

From that, we can see that four results are wrong for CCA. ACMR and our method have obtained good retrieval performance in the task of Img2Txt. For the task of Txt2Img, the CCA retrieved some wrong images, which do not belong to geography category. For ACMR, there is a wrong result. Compared to ACMR, the results retrieved by

**Fig. 6** Running time of ReSAN and ACMR

ReSAN are more related to the semantic category although there is also a wrong result. The reason might be that CCA measures the global correlation between data from different modalities. However, our method and ACMR explore the semantic information of each modality data based deep convolutional network, which can reduce the semantic gap effectively and generate more discriminative representations.

Furthermore, we visualize the learned representations on the Wikipedia dataset for ACMR and ReSAN using t-SNE [42] in Fig. 4. We apply min–max normalization to make the distribution of each category to be more clear. From that, we conclude the learned representation by our method are more discriminative compared to other method. Specifically, we can see from the figure that the representation generated for images on some classes (biology, geography, sport, warfare) are relatively concentrated. Because our method separate the original representation into common representation and private representation. Compared with the original representation, the common representation obtains more modality independent semantic information which is more helpful to reduce the modality gap. This demonstrates the effectiveness of the representation separation.

5.3.2 Results on NUSWIDE-10k dataset

The results on NUSWIDE-10k are presented in Table 3. From this Table, we obtain following observations: (1) JRL achieved best performance among the traditional methods, and demonstrate the advantages of jointly using supervised information and graph regularization, (2) Among the DNN-based methods, ACMR and our method obtained the best performances.

Figure 5 shows the PR curves of ACMR and ReSAN on Wikipedia and NUS-WIDE-10k dataset. From that, we can see the PR of ReSAN is better than that of ACMR on Wikipedia dataset. For NUS-WIDE-10k dataset, ReSAN achieved comparable results with ACMR.

5.3.3 The effectiveness of different terms

In the proposed method, adversarial learning aims to model the joint data distribution of different modalities, while representation separation is to learn common semantic representations for cross-modal retrieval. To demonstrate their contribution for improving the retrieval performance, we denote the ReSAN as three different model, without adversarial learning as (ReSAN-D), without representation separation as (ReSAN-P), without reconstruction as (ReSAN-C). The results are shown in the Table 4. From that, we can see that three different components effectively improve the retrieval performance with different levels.

Among them, the representation separation and reconstruction plays important roles for the retrieval performance.

5.3.4 Running time of ReSAN

Figure 6 shows the running time of ReSAN and ACMR. It can be seen that our method takes longer time to compute representations than that of ACMR. In fact, it can be done off-line. During the retrieval stage, our method is faster than ACMR. This is very important in real applications.

6 Conclusion

In this paper, we proposed a representation separation adversarial networks for cross-modal retrieval method, which explicitly splits the original representations into common representation and private representation for each modality. To learn modality-variant common representation, we proposed exchanging strategy the common representations among different modalities. Furthermore, we adopt the label information to increase the discriminant ability for the common representations. The experimental results on two wide datasets demonstrate as follows: modeling the unique part for each modality can effectively improve the robustness of the common representations. In the future, we will apply this representation separation approach for the unsupervised scenarios.

Acknowledgements Weihua Ou and Heping Song are the corresponding authors. This work was supported by the National Natural Science Foundation of China (Nos. 61762021, 61962010, 61976107), Natural Science Foundation of Guizhou Province (Grant Nos. [2017]1130, [2017]5726-32), Excellent Young Scientific and Technological Talent of Guizhou Province ([2019]-5670), Natural Science Foundation of Jiangsu Province under Grant BK20170558.

References

- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).
- Lu, H., Zhang, M., Xu, X., Li, Y., & Shen, H. T. (2020). Deep fuzzy hashing network for efficient image retrieval. *IEEE Transactions on Fuzzy Systems*. <https://doi.org/10.1109/TFUZZ.2020.2984991>.
- Lu, W., Zhang, X., Lu, H., & Li, F. (2020). Deep hierarchical encoding model for sentence semantic matching. *Journal of Visual Communication and Image Representation*. <https://doi.org/10.1016/j.jvcir.2020.102794>.
- Zhang, Y., Lu, W., Ou, W., Zhang, G., Zhang, X., Cheng, J., & Zhang, W. (2019). Chinese medical question answer selection via hybrid models based on CNN and GRU. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-019-7240-1>.
- Peng, L., Yang, Y., Ji, Y., Lu, H., & Shen, H. T. (2019). Coarse to fine: Improving VQA with cascaded-answering model. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2019.2903516>.
- Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 154–162).
- Peng, Y., Huang, X., & Zhao, Y. (2017). An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9), 2372–2385.
- Xu, X., He, L., Lu, H., Gao, L., & Ji, Y. (2019). Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2), 657–672.
- Zhang, J., & Peng, Y. (2019). Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(1), 174–187.
- Xu, X., Shen, F., Yang, Y., Shen, H. T., & Li, X. (2017). Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5), 2494–2507.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3–4), 321–377.
- Rasiwasia, N., Pereira, J. C., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 251–260).
- Akaho, S. (2006). A kernel method for canonical correlation analysis. arXiv preprint [arXiv:cs/0609071](https://arxiv.org/abs/0609071).
- Wang, W., & Livescu, K. (2015). Large-scale approximate kernel canonical correlation analysis. arXiv preprint [arXiv:1511.04773](https://arxiv.org/abs/1511.04773).
- Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3441–3450).
- Feng, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *The 22nd international conference on multimedia (ACM)* (pp. 7–16).
- Peng, Y., Huang, X., & Qi, J. (2016). Cross-media shared representation by hierarchical learning with multiple deep networks. In *IJCAI* (pp. 3846–3853).
- Peng, Y., & Qi, J. (2019). CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1), 1–24.
- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247–1255).
- Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation learning. In *International conference on machine learning* (pp. 1083–1092).
- Jiang, Q.-Y., & Li, W.-J. (2017). Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3232–3240).
- Li, C., Deng, C., Li, N., Liu, W., Gao, X., & Tao, D. (2018). Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4242–4251).
- Yang, E., Deng, C., Li, C., Liu, W., Li, J., & Tao, D. (2018). Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5292–5303.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).
- Zhang, Y., & Lu, H. (2018). Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 686–701).

26. Wang, X., & Gupta, A. (2016). Generative image modeling using style and structure adversarial networks. In *European conference on computer vision* (pp. 318–335). Springer.
27. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint [arXiv:1605.05396](https://arxiv.org/abs/1605.05396).
28. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).
29. Yan, B., Wang, D., Lu, H., & Yang, X. (2020). Cooling-shrinking attack: Blinding the tracker with imperceptible noises. arXiv preprint [arXiv:2003.09595](https://arxiv.org/abs/2003.09595).
30. Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5542–5551).
31. Tong, B., Klinkigt, M., Chen, J., Cui, X., Kong, Q., Murakami, T., & Kobayashi, Y. (2018). Adversarial zero-shot learning with semantic augmentation. In *Thirty-second AAAI conference on artificial intelligence*.
32. Wu, L., Wang, Y., & Shao, L. (2018). Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4), 1602–1612.
33. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain separation networks. In *Advances in neural information processing systems* (pp. 343–351).
34. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).
35. Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R. G., Levy, R., et al. (2013). On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3), 521–535.
36. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: A real-world web image database from national university of Singapore. In *Proceedings of the ACM international conference on image and video retrieval* (pp. 1–9).
37. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
38. Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12), 2639–2664.
39. Wang, K., He, R., Wang, W., Wang, L., & Tan, T. (2013). Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE international conference on computer vision* (pp. 2088–2095).
40. Zhai, X., Peng, Y., & Xiao, J. (2014). Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6), 965–978.
41. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *The 28th international conference on machine learning (ICML)* (pp. 689–696).
42. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Jiaxin Deng is current an Master Degree Candidate at Guizhou Normal University, Guiyang, China. His research interests include multimedia information retrieval and deep learning.



Weihua Ou received the M.S. degree in Mathematics from the Southeast University, Nanjing, China in 2006 and the Ph.D. degree in Information and Communication Engineering from Huazhong University of Science and Technology (HUST), China in 2014, respectively. Currently, he is an full Professor at the School of Big data and Computer Science in Guizhou Normal University, Guiyang, China. His current research interests include cross-

modal retrieval, deep learning, and image processing and computer vision.



Jianping Gou received the Ph.D. degree in computer science from University of Electronic Science and Technology of China, Chengdu, China, in 2012. From 2018 to 2020, he was a Postdoctoral Fellow at UBTECH Sydney AI Centre, The University of Sydney. He is currently an Associate Professor in School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, China. His current research interests

include pattern classification, machine learning.



Heping Song received the Ph.D. degree in computer application technology from Sun Yat-sen University, Guangzhou, China, in 2011. He is currently a Associate Professor with the Department of Software Engineering, School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His current research interests include low-level vision, inverse problems, computer vision and deep learning.



Anzhi Wang received the Ph.D. degree from Sichuan University in 2017, supervised by Prof. Ming-Hui Wang. He is currently an Lecturer at School of Big Data and Computer Science, Guizhou Normal University. His research interests includes computer vision, image processing, and machine learning. He reviews papers regularly for journals including the TMM, IEEE-SPL, IET-CV, IET-IP, NPL, and MTA.



Xing Xu received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015. He is currently with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests mainly focus on multimedia information retrieval, pattern recognition, and computer vision.