Modality Consistent Generative Adversarial Network for Cross-Modal Retrieval

Zhiyong Wu[1], Fei Wu[1(✉)], Xiaokai Luo[1], Xiwei Dong[1], Cailing Wang[1], and Xiao-Yuan Jing[2]

[1] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China
wuzybarskish@163.com, wufei_8888@126.com, wangcl@njupt.edu.cn

[2] School of Computer, Wuhan University, Wuhan, China
jingxy_2000@126.com

Abstract. Cross-modal retrieval, which aims to perform the retrieval task across different modalities of data, is a hot topic. Since different modalities of data have inconsistent distributions, how to reduce the gap of different modalities is the core of cross-modal retrieval issue. Recently, Generative Adversarial Networks has been used in cross-modal retrieval due to its strong ability to model data distribution. We propose a novel approach named Modality Consistent Generative Adversarial Network for cross-modal retrieval (MCGAN). The network integrates a generator to generate synthetic image features from text features, a discriminator to classify the modality of features, and followed by a modality consistent embedding network that projects the generated image features and real image features into a common space for learning the discriminative representations. Experiments on two datasets prove the performance of MCGAN on cross-modal retrieval, compared with state-of-the-art related works.

Keywords: Generative adversarial network · Cross-modal retrieval

1 Introduction

Nowadays, a large amount of multimedia data with different modalities, e.g., image, text, video, etc., is mixed together to gain a comprehensive understanding of the real world. The existence of the huge multi-modal data repository greatly stimulates the demand for cross-modal retrieval in search engines or digital libraries, such as returning concerned results from image as response to query of text or vice versa. Cross-modal retrieval provides queries against any modality to find relevant information with different modalities [1].

The main task of cross-modal retrieval is to bridge the modality gap. A large body of traditional cross-modal retrieval methods have been proposed to learn linear projections by optimizing the statistical values from different modalities into a common semantic space and explore the correlation, like canonical correlation analysis

The first author is a student.

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11859, pp. 3–14, 2019.
https://doi.org/10.1007/978-3-030-31726-3_1

(CCA)-based methods [2]. Deep learning technology is widely used in image recognition, natural language processing and object dictation [3, 4]. Deep neural network can also play a good role in the field of cross-modal retrieval. Deep neural network (DNN)-based methods [5–7] construct multilayer network to conduct nonlinear projection. The correlation learning error across different modalities is minimized for bridging the gap of different modalities and learning the common representation. Recent works have shown that generative adversarial networks (GANs) [8] have the advantage of modeling data distribution. Inspired by GANs, the heterogeneous gap of different modalities can be reduced through the adversarial mechanism, and some GAN-based cross-modal retrieval methods are proposed [9, 10].

1.1 Motivation and Contribution

Although many methods were proposed focusing on cross-modal retrieval research, how to better bridge the gap of different modalities and improve the accuracy of retrieval are still concerned [11]. Most of existing methods [12, 13] project data from different modalities into a common semantic space in which the similarity measurements are made. However, these methods directly project data from different modalities into common semantic space to reduce the gap, which will lead to the loss of semantic information in both image and text modalities. How to effectively reduce the heterogeneous gap and retain the semantic information of each modality as much as possible has not been well studied.

Inspired by [14] that leverages GANs as a powerful model to convert cross-modal data to single-modal data for zero-shot learning, we propose a novel approach named Modality Consistent Generation Adversarial Network for cross-modal retrieval (MCGAN). The contributions of our study are three-fold:

(1) We design a new generative adversarial network to generate image features with the input text features, which projects text features into the image feature space. In this way, the cross-modal retrieval problem is converted into a single-modal retrieval problem. The gap of different modalities is bridged while the image semantic information is preserved as much as possible.
(2) We project the generated image features and real image features into a common space via a sub-network, and utilize label information to model both the inter- and intra-modal similarity, such that features are semantically discriminative in both inter- and intra-modal aspects.
(3) MCGAN is evaluated on two widely used datasets, i.e., Wikipedia dataset [2] and NUS-WIDE-10 k dataset [5]. The experimental results show that it can outperform related state-of-the-art works.

2 Related Work

2.1 Non-GANs-Based Cross-Modal Retrieval Methods

There exist many methods proposed to bridge the heterogeneity gap between different modalities, which focus on learning common representation of different modalities and measuring similarities to correlate the heterogeneous data [15].

Traditional cross-modal retrieval methods usually linearly project cross-modal data into a common space to generate the common representation. The similarity measurement of features in the common space can maximize the correlation between modalities. Based on canonical correlation analysis (CCA) [16], some representative methods are developed for cross-modal retrieval. Rasiwasia et al. [2] project text features and image features into a low-dimensional common subspace and investigate the correlation between two modalities through CCA. After [2], plenty of extensions, for example [17] adopts kernel function to pursue features and incorporate the semantic labels for learning correlation between two modalities. Besides, Wang et al. [12] present a method learning coupled feature spaces (LCFS) to learn a coupled feature space by coupled linear regression, and the selection of discriminant and relevant features is considered in the space. Furthermore, joint feature selection and subspace learning (JFSSL) [18] method integrates graph regularization and label information to make inter- and intra-modalities features close to relevant labels while far away from irrelevant labels.

Recently, deep neural network promotes the development of cross-modal retrieval due to its great nonlinear fitting ability and self-learning ability [19]. Deep learning based methods non-linearly project the data of each modality to independent semantic space for feature extraction. Feng et al. [5] propose correspondence autoencoder (Corr-AE), which takes representation learning and correlation learning into account to establish a robust model. Since the convolutional neural network (CNN) can fit the image well to get the visual features, Wei et al. [13] provide a deep semantic matching (Deep-SM), which adopts CNN to get deep visual features, validating the superiority of CNN for improving the performance of cross-modal retrieval. Cross-media multiple deep network (CMDN) presented by Peng et al. [6] obtains separate representation of each media type through a model that combines intra- and inter-media representations hierarchically to get the shared representations.

2.2 Generative Adversarial Networks (GANs)-Based Cross-Modal Retrieval Methods

Generative Adversarial Networks proposed by Goodfellow et al. [8] is an unsupervised learning model, which is used to generate desired image from random noise. After several years of development, it has been used in many applications, such as image style transformation, object detection, zero-shot learning including cross-modal retrieval. The original GANs can been divided into two models: generator G and discriminator D. The two models carry out alternating iterative training in the way of minimax game and finally enable generator G to learn the data distribution of real images. Generator G receives random noise, obtains the distribution of real images and outputs the generated images, while the discriminator D aims to distinguish whether the input image is real or not.

However, the original GANs has the problems of unstable training, gradient disappearance and mode collapse, which makes the generated results unsatisfactory. In order to solve these problems, Arjovsky et al. [20] put forward Wasserstein GAN training strategy and adopt gradient penalty to train the model. Condition generative adversarial networks (CGANs) [21] is proposed to add constraint conditions for GANs.

The data is labeled in the generative model and discriminative model respectively, so as to increase the clarity of the images generated by GANs. Radford et al. propose deep convolutional generative adversarial networks (DCGAN) [22], which applies the convolutional neural network to GANs to make the generated images more precise. Recently, Wang et al. [10] apply GANs to cross-modal retrieval and propose adversarial cross-modal retrieval (ACMR), which projects features of different modalities into common space through the minimax training strategy to obtain discriminative feature representations.

These non-GANs-based methods and GANs-based methods directly project data of different modalities into common semantic space to reduce the gap, which will lead to the loss of semantic information in both image and text domain. Different from them, our method effectively transforms the cross-modal retrieval issue into the single-mode retrieval issue, and retains the semantic information of each modality while reducing the heterogeneous gap of different modalities. In addition, we use the label information to model the similarity between and within modality, and obtain semantically more discriminative feature representations.

3 Our Approach

3.1 Problem Formulation

Let $\Omega = \{o_n, y_n\}_{n=1}^{N}$ be a set of N instances of paired image and text, where each instance $o_n = (v_n, t_n)$ includes an image feature vector $v_n \in \mathbb{R}^{d_v}$ and a text feature vector $t_n \in \mathbb{R}^{d_t}$, d_v and d_t denote the feature dimension of two modalities, and n is the number of training samples. Let $V = [v_1, \ldots, v_N]$ and $T = [t_1, \ldots, t_N]$ be the training sets of image features and text features, respectively. $y_n = [y_{n1}, \ldots, y_{nC}]^T$ denotes the semantic category label vector corresponding to o_n, where $y_{nc} = 1$ if $o_n = (v_n, t_n)$ is from the c^{th} class while $y_{nc} = 0$ otherwise. The generator G is designed to learn synthetic image feature representations $\tilde{V} = G(T; \theta) = [\tilde{v}_n]_{n=1}^{N} \in \mathbb{R}^{d_v \times N}$ for text modality. To explore the correlation between modalities, we adopt a common two layers feed-forward sub-networks to nonlinearly project V and $\tilde{V}$ into a common space for learning the correlative representations, by $V_s = f(V; \phi) = \left[s_v^n\right]_{n=1}^{N} \in \mathbb{R}^{d_s \times N}$ and $\tilde{V}_s = f\left(\tilde{V}; \phi\right) = \left[s_{\tilde{v}}^n\right]_{n=1}^{N} \in \mathbb{R}^{d_s \times N}$, where $f(\cdot; \phi)$ is the mapping function.

The objectives of our approach can be summarized as two points: (1) the text features can be effectively converted into the space of image features through adversarial mechanism; (2) the learned features should be semantically discriminative. We alternately and iteratively train the generator G, discriminator D and common embedding network respectively. Figure 1 shows the overall framework of MCGAN.

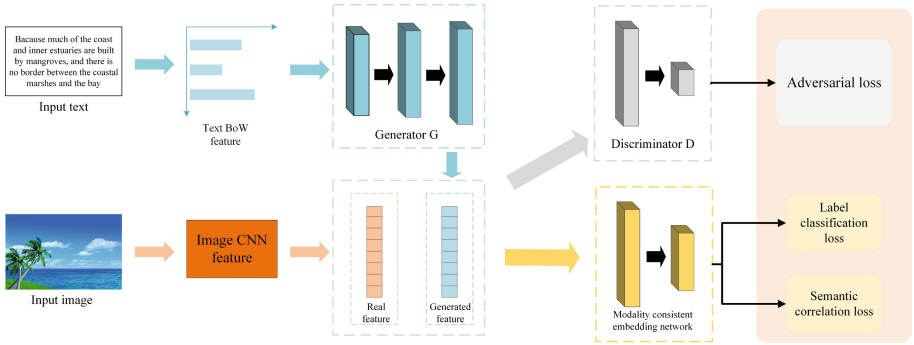

Fig. 1. Our MCGAN overall framework. The architecture of MCGAN consists of two parts. (1) A generative adversarial network is composed of generative model G and discriminative model D: the generative model G takes the text features as input and outputs the generated features near to the real image features; the discriminative model tries to distinguish the real and generated image features via the adversarial loss. (2) A modality consistent embedding network is a two feed-forward sub-network, which models both the intra-modal semantic similarity via label classification loss and the inter-modal semantic similarity via semantic correlation loss.

3.2 Generative Model

Our generative adversarial networks (GANs) defines a minimax game between two competing components: a generator G that captures the image feature distributions from text features for synthesizing image features, and a discriminator D that is learned to distinguish the real image features from synthetic features. Specifically, text features T which are extracted by a well-known bag-of-words (BoW) vector with the TF-IDF weighting scheme, are accepted as input by three-layer feed-forward networks, and the generated image features $\tilde{V} = G(T;\theta)$. In the minimax game, the goal of generator G is to make the synthetic image features approximate to the real image features through the adversarial training strategy. Inspired by Wasserstein GAN that is stable to synthesize great images, the loss of generator is defined as:

$$L_G = -E_{T\sim p_T}[D(G(T;\theta);\omega)] \tag{1}$$

where θ and ω denote the parameters of generator and discriminator respectively, p_T is the distribution of text features.

3.3 Discrimination Model

The discriminator D is actually a modality classifier used to distinguish whether the input features are real image features or not. In generative adversarial networks, the discriminator D plays the role of adversary, distinguishing input feature by minimizing the classification error of probabilities $D(V;\omega)$ and $D(\tilde{V};\omega)$. As shown in Fig. 1, we build a modality classifier with a two-layer sub-network, which takes as input either a real image feature or a generated image feature and the outputs are $D(V;\omega)$ and

$D(\tilde{V};\omega)$. In other to solve the problems of unstable training and mode collapse of GAN, the training strategy of Wasserstein GAN is adopted to train the discriminator through calculating the Wasserstein distance of the distribution of real image features and synthetic image features as loss. Furthermore, a differentiable Lipschitz constraint with gradient penalty is added to prevent the gradient from disappearing during training. The loss for the discriminator is formulated as:

$$\begin{aligned} L_D = & E_{T\sim p_T}[D(G(T;\theta);\omega)] - E_{v\sim p_v}[D(V;\omega)] \\ & + \lambda E_{\hat{V}\sim p_{\hat{V}}}\left[\left(\left\|\nabla_{\hat{v}} D(\hat{V};\omega)\right\|_2 - 1\right)^2\right] \end{aligned} \tag{2}$$

where $\hat{V}$ is the linear interpolation of real image feature V and generated image feature $\tilde{V}$. The first two terms approximate Wasserstein distance of distribution of real image feature V and generated image feature $\tilde{V}$. The third term is the gradient penalty to enforce the Lipschitz constraint with λ being the penalty coefficient.

3.4 Modality Consistent Embedding Network

Though we have obtained the distribution of image features through the generative adversarial network and have converted the cross-modal retrieval issue into single-modal retrieval issue, the similarity measurement of paired features is also what we should focus on. In other to capture more discriminative features semantically, we propose a modality consistent embedding network, which is a two-layer sub-network, mapping paired features into a common space, and then label information is used to model the inter- and intra-modal semantic similarity.

Intra-modal Semantic Similarity Modeling
To make the paired features to be semantically discriminative, a feed-forward one-layer sub-network activated by Softmax is adopted as a classifier, such that when the output of feature embedding network $s_{v_n} = f(v_n;\phi)$ or $s_{\tilde{v}_n} = f(\tilde{v}_n;\phi)$ is the input of classifier, the corresponding probability distribution of semantic categories, i.e., $\hat{p}_n(s_{v_n})$ or $\hat{p}_n(s_{\tilde{v}_n})$ can be output. We define the following label classification loss:

$$L_C = -\frac{1}{N}\sum_{n=1}^{N} y_n(\log \hat{p}_n(s_{v_n}) + \log \hat{p}_n(s_{\tilde{v}_n})) \tag{3}$$

where y_n is the ground-truth label of each feature, which is expressed as an one-hot vector.

Inter-modal Semantic Similarity Modeling
The embedding features of two modalities in the common space have superior intra-modal semantic similarity through the combination of GAN and feature embedding network. Furthermore, in order to get better classification results, the embedding features should also show good inter-modal semantic similarity. Motivated by [14], we design a modality consistent semantic correlation term to calculate the similarity of

features with the same semantic category. We provided the following semantic correlation loss:

$$L_m = \frac{1}{C}\sum_{c=1}^{C}\left\| E_{s_{\tilde{v}_c} \sim p_{\tilde{v}}^c}[s_{\tilde{v}_c}] - E_{s_{v_c} \sim p_v^c}[s_{v_c}] \right\|^2 \tag{4}$$

where C is the number of classes, s_{v_c} is the embedding image feature of class c and $s_{\tilde{v}_c}$ is the embedding generated feature of class c. For each modality, the centroid of the cluster of embedding features should be defined, so we adopt the empirical expectation $E_{x_c \sim p_x^c}[x_c]$ to calculate the centroid of the embedding features of class c. We define the following formulas as:

$$\begin{aligned} E_{s_{v_c} \sim p_v^c}[s_{v_c}] &= \frac{1}{U_c}\sum_{i=1}^{U_c} s_{v_c}^i \\ E_{s_{\tilde{v}_c} \sim p_{\tilde{v}}^c}[s_{\tilde{v}_c}] &= \frac{1}{M_c}\sum_{i=1}^{M_c} s_{\tilde{v}_c}^i \end{aligned} \tag{5}$$

where the first formula is the expectation of embedding image features, which is approximated by averaging the embedding image features for class c, and U_c is the number of samples in class c. Similarity, the second formula is the expectation of embedding generated features, and M_c is the number of embedding generated features for class c.

By combining the Eqs. (4) and (5), we obtain the optimization loss of the modality consistent embedding network for learning discriminative features as follows

$$L_{emb} = L_C + \zeta L_m \tag{6}$$

where ζ is a parameter to balance two terms.

3.5 Optimization

The overall framework proposed in this paper is composed of two components: a generative adversarial network to generate the generated image features that are close to real image features, and a modality consistent embedding network to obtain more discriminative features. The optimal features can be obtained by integrating the loss functions in Eqs. (2), (3) and (7). The optimization problems for discriminator D, generator G and modality consistent embedding network are respectively defined as follows:

$$\left(\widehat{\omega}\right) = \arg\min_{\omega}(L_{emb} + \alpha L_D) \tag{7}$$

$$\left(\widehat{\theta}\right) = \arg\min_{\theta}(L_{emb} + \beta L_G) \tag{8}$$

$$\left(\hat{\phi}\right) = \arg\min_{\phi}(L_{emb}) \tag{9}$$

where α and β are tradeoff parameters. Each part of the network is updated separately though the optimization objectives above. The parameters ω, θ and ϕ can be effectively optimized through the automatic differential back propagation of Pytorch. Algorithm 1 summarizes the process of our approach.

Algorithm 1 Optimization procedure of MCGAN

1. **Input:** mini-batch image features $V = [v_1, ..., v_N]$ and text features $T = [t_1, ..., t_N]$, the semantic category label $y_n = [y_{n1}, ..., y_{nC}]^T$ and number of training epoch S.
2. **Training procedure:**
   (1) Initialize generative network G, discriminative network D and modality consistent embedding network;
   (2) **for** $i = 1$ to S **do**
   $\tilde{V} \leftarrow G(T;\theta)$; $s_{v_n} = f(v_n;\phi)$; $s_{\tilde{v}_n} = f(\tilde{v}_n;\phi)$;
   Compute $L_{emb} + \alpha L_D$ using Eqs. (2) and (6);
   Update ω by $Adam(\nabla_{\omega} L_{emb} + \alpha L_D)$;
   Compute $L_{emb} + \beta L_G$ using Eqs. (1) and (6);
   Update θ by $Adam(\nabla_{\theta} L_{emb} + \beta L_G)$;
   Compute L_{emb} using Equation (6);
   Update ϕ by $Adam(\nabla_{\phi} L_{emb})$;
   end for
3. **Output:** Optimized parameters ω, θ, ϕ.

4 Experiments

4.1 Datasets

We evaluate our proposed approach on the widely used Wikipedia dataset [2] and NUS-WIDE-10 k dataset [5].

Wikipedia dataset is collected from Wikipedia featured articles, and there are 2,866 image-text pairs. Each pair of image and text is extracted from the same articles. All image-text pairs are from 10 semantic classes, and each pair is labeled with only one class label. Following [2], 2,173 pairs of samples are used for training, 231 pairs for validation and 462 pairs for testing. For image modality, 4,096-dimensional features are extracted by fc7 layer of VGGnet, and each text is represented by a 3,000-demensional Bag-of-Word feature.

NUS-WIDE-10 k dataset consists of 10,000 web images including 10 semantic concepts download from Flicker website. Following [5], this dataset is split into three subsets: the training set with 8,000 pairs, the validation set with 1,000 pairs, and the testing set with 1,000 pairs. For each image, 4,096-dimensional feature is extracted by the fc7 layer of VGGnet, and for each text, 1,000-dimensional Bag-of-Word feature vector is extracted.

4.2 Evaluation Measure and Compared Methods

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

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP(q_i) \tag{10}$$

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.

For comparison, we compare our proposed MCGAN approach with six representative cross-modal retrieval methods: (1) traditional cross-modal retrieval methods: CCA [2] and LCFS [12]; (2) deep learning-based methods: Deep-SM [13], Corr-AE [5] and MCSM [23]; (3) GAN-based method: ACMR [10]. We report the experimental results of the compared methods according to the published results in their papers or the codes provided by the authors to implement the evaluation.

In experiment, we perform two types of experiments, namely retrieving images with text and retrieving text with images.

4.3 Implementation Detail

Our proposed MCGAN approach and relevant experiments are implemented on Torch framework. The implementation details of our generative adversarial network and the modality consistent embedding network are as follows: our generative adversarial network consists of two components, the generative model is a 3-layer network, which is composed of three fully connected layers to learn the generated image features from text features. The number of neurons in each layer is 3500, 4000, 4096, and the activation function is Tanh. The discriminative model consists of two fully connected layers: the number of neurons in the first layer is 1000, the number of neurons in the second layer is 2, and the subsequent activation function is ReLU. In addition, Softmax activation is added after the last layer to conduct the modality classification. For the modality consistent embedding network, two fully connected layers with dimensional [1000, 10] activated by Tanh are used to project both the generated image features and the real image features into a common semantic space to learn the discriminative feature representations.

In our training procedure, the mini-batch size is 128, and the tradeoff parameters λ, ζ, α and β are set up by grid search. The good results are achieved with $\lambda = 10$, $\zeta = 1$, $\alpha = \beta = 0.1$.

4.4 Result and Discussion

Table 1 tabulates the Map results of compared methods on Wikipedia and NUS-WIDE-10 k datasets. In can be seen from the table that in both image to text and text to image retrieval tasks, GAN-based methods such as MCSM and ACMR outperform the non-GAN-based cross-modal retrieval methods including CCA, LCFS, Corr-AE, CMDN and Deep-SM on the benchmark datasets. Furthermore, our MCGAN can always outperform all compared methods. Specifically, for the retrieval task of image to text and text to image on the Wikipedia dataset, MCGAN improves the mAP results at least by 0.004 = (0.522 − 0.518), 0.013 = (0.471 − 0.458). Similarly, for the retrieval task of image to text and text to image on the NUS-WIDE-10 k dataset, our approach improves the mAP scores at by 0.019 = (0.563 − 0.544), 0.01 = (0.551 − 0.541). The results show that by turning text features into image features through generative adversarial network, semantic information can be effectively preserved while the gap of different modalities can be bridged. Besides, the more discriminative features learned from the inter- and intra-modal discrimination are helpful to improve the retrieval performance.

Table 1. The mAP cross-modal retrieval results on two datasets

Method	Wikipedia			NUS-WIDE-10 k		
	Img2txt	Txt2img	Average	Img2txt	Txt2img	Average
CCA	0.258	0.250	0.254	0.202	0.220	0.211
LCFS	0.455	0.398	0.427	0.383	0.346	0.365
Corr-AE	0.402	0.395	0.399	0.366	0.417	0.392
CMDN	0.488	0.427	0.458	0.492	0.515	0.504
Deep-SM	0.458	0.345	0.402	0.389	0.496	0.443
MCSM	0.516	0.458	0.487	0.543	0.541	0.542
ACMR	0.518	0.412	0.465	0.544	0.538	0.541
MCGAN	**0.522**	**0.471**	**0.497**	**0.563**	**0.551**	**0.557**

In the modality consistent embedding network, label classification loss and semantic correlation loss are defined to promote semantically discriminative feature learning. To demonstrate whether they can contribute to improving the retrieval performance, the version of MCGAN without label classification loss (MCGAN-C), the version of MCGAN without semantic correlation loss (MCGAN-m) are proposed to evaluate the role of each component. From the Table 2, the mAP results of MCGAN-C, MCGAN-m and MCGAN show that the label classification loss and semantic correlation loss are contributed to promoting semantically discriminative feature learning and improve the retrieval performance.

Table 2. The mAP results of cross-modal retrieval with fully MCGAN, MCGAN without L_C, and MCGAN without L_m.

Method	Wikipedia			NUS-WIDE-10 k		
	Img2txt	Txt2img	Average	Img2txt	Txt2img	Average
MCGAN	**0.522**	**0.471**	**0.497**	**0.563**	**0.551**	**0.557**
MCGAN-C	0.234	0.188	0.211	0.203	0.181	0.192
MCGAN-m	0.480	0.422	0.451	0.511	0.506	0.508

5 Conclusion

In this paper, we present a novel approach named MCGAN that is able to convert the cross-modal retrieval issue into single-modal retrieval issue on image domain via generative adversarial network. In this way, the semantic information of image modality can be preserved effectively. Furthermore, a modality consistent embedding network is designed to project both the image features and generated image features to a common semantic space and utilize label information to model both the inter- and intra-modal similarity via two defined loss functions. Extensive empirical results demonstrate that MCGAN can achieve significantly better retrieval performance than several state-of-the-art related methods.

Acknowledgements. The work in this paper was supported by National Natural Science Foundation of China (No. 61702280), Natural Science Foundation of Jiangsu Province (No. BK20170900), National Postdoctoral Program for Innovative Talents (No. BX20180146), Scientific Research Starting Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX17_0794.

References

1. Li, D., Dimitrova, N., Li, M., et al.: Multimedia content processing through cross-modal association. In: ACM International Conference on Multimedia, Berkeley, 2–8, pp. 604–611 (2003)
2. Rasiwasia, N., Costa Pereira, J., Coviello, E., et al.: A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia, pp. 251–260 (2010)
3. Fan, D.P., Wang, W., Cheng, M.M., et al.: Shifting more attention to video salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8554–8564 (2019)
4. Zhao, J.X., Cao, Y., Fan, D.P., et al.: Contrast prior and fluid pyramid integration for RGBD salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
5. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM International Conference on Multimedia, pp. 7–16 (2014)
6. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: International Joint Conference on Artificial Intelligence, pp. 3846–3853 (2016)

7. Huang, X., Peng, Y., Yuan, M.: Cross-modal common representation learning by hybrid transfer network. In: International Joint Conference on Artificial Intelligence, pp. 1893–1900 (2017)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
9. Peng, Y., Qi, J.: CM-GANs: cross-modal generative adversarial networks for common representation learning. ACM Trans. Multimed. Comput. Commun. Appl. **15**(1), 22 (2019)
10. Wang, B., Yang, Y., Xu, X., et al.: Adversarial cross-modal retrieval. In: ACM International Conference on Multimedia, pp. 154–162 (2017)
11. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450 (2015)
12. Wang, K., He, R., Wang, W., et al.: Learning coupled feature spaces for cross-modal matching. In: IEEE International Conference on Computer Vision, pp. 2088–2095 (2013)
13. Wei, Y., Zhao, Y., Lu, C., et al.: Cross-modal retrieval with CNN visual features: a new baseline. IEEE Trans. Cybern. **47**(2), 449–460 (2017)
14. Zhu, Y., Elhoseiny, M., Liu, B., et al.: A generative adversarial approach for zero-shot learning from noisy texts. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1004–1013 (2018)
15. Wu, F., et al.: Cross-project and within-project semisupervised software defect prediction: a unified approach. IEEE Trans. Reliab. **67**(2), 581–597 (2018)
16. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
17. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: IEEE International Conference on Computer Vision (2015)
18. Wang, K., He, R., Wang, L., et al.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2010–2023 (2016)
19. Wu, F., et al.: Intraspectrum discrimination and interspectrum correlation analysis deep network for multispectral face recognition. IEEE Trans. Cybern. 1–14 (2018)
20. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223 (2017)
21. Isola, P., Zhu, J.Y., Zhou, T., et al.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
22. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
23. Peng, Y., Qi, J., Yuan, Y.: Modality-specific cross-modal similarity measurement with recurrent attention network. IEEE Trans. Image Process. **27**(11), 5585–5599 (2018)