Neurocomputing 362 (2019) 118–128

Contents lists available at ScienceDirect

Neurocomputing

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

Adversarial task-specific learning

Xin Fu[a,b], Yao Zhao[a,b,*], Ting Liu[a,b], Yunchao Wei[c], Jianan Li[d], Shikui Wei[a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
[c] Beckman Institute, University of Illinois Urbana-Champaign, the United States
[d] School of Optical Engineering, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history:
Received 27 January 2019
Revised 20 May 2019
Accepted 15 June 2019
Available online 20 July 2019

Communicated by Dr. Iosifidis Alexandros

Keywords:
Cross-modal retrieval
Adversarial learning
Subspace learning

ABSTRACT

In this paper, we investigate a principle way to learn a common feature space for data of different modalities (*e.g.* image and text), so that the similarity between different modal items can be directly measured for benefiting cross-modal retrieval task. To effectively keep semantic/distribution consistent for common feature embeddings, we propose a new Adversarial Task-Specific Learning (ATSL) approach to learn distinct embeddings for different retrieval tasks, *i.e.* images retrieve texts (I2T) or texts retrieve images (T2I). In particular, the proposed ATSL is with the following advantages: (a) semantic attributes are leveraged to encourage the learned common feature embeddings of couples to be semantic consistent; (b) adversarial learning is applied to relieve the inconsistent distribution of common feature embeddings for different modalities; (c) triplet optimization is employed to guarantee that similar items from different modalities are with smaller distances in the learned common space compared with the dissimilar ones; (d) task-specific learning produces better optimized common feature embeddings for different retrieval tasks. Our ATSL is embedded in a deep neural network, which can be learned in an end-to-end manner. We conduct extensive experiments on two popular benchmark datasets, *e.g.* Flickr30K and MS COCO. We achieve R@1 accuracy of 57.1% and 38.4% for I2T and 56.5% and 38.6% T2I on MS COCO and Flickr30K respectively, which are the new state-of-the-arts.

1. Introduction

The prevalence of social networking has significantly increased the amount of data with various modalities, and images and texts contribute the dominant forms of the data. Usually, the data with different modalities are leveraged collectively to describe the same object or event. For instance, an image generally carries similar semantic information with texts. Consequently, the cross-model retrieval task is imperative for mining the association between different modalities. In this work, we focus mainly on tackling the cross-model retrieval task. However, cross-modal data usually spans different feature spaces and has its own distribution characteristics. Thus, measuring the semantic similarity between visual data and text data has been widely considered as a great challenge for cross-modal learning communities.

In the last few decades, numerous methods [1–6] have been proposed for the cross-model retrieval. Most existing approaches try to learn an optimal embedding for each view or modality in a common latent space, in which the distance between different modal items with similar semantics could be minimized while that with dissimilar semantics could be maximized. In the common space, the items from different modalities have isomorphic representations so that the similarity between them can be directly computed. We observe that most previous works have learned only one common space for addressing both images retrieve texts (I2T) and texts retrieve images (T2I) tasks. However, such a common space may not be optimal one for I2T or T2I, which has already revealed by Wei et al. [7]. We consider the reason as the learned common space is a compromised one by taking both two retrieval constraints into account. Although Wei et al. [7] have leveraged two independent common spaces for addressing different retrieval tasks, there are some obvious shortages: (1) it is heavily dependent on high-level data categories when learning the common representations; (2) it only considers to correlate common embeddings using pair-wise inputs but ignores the dissimilar samples; (3) consistent distribution of the learned embedding features for both images and texts is not taken into account.

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
E-mail addresses: xinfu@bjtu.edu.cn (X. Fu), yzhao@bjtu.edu.cn (Y. Zhao), 16112055@bjtu.edu.cn (T. Liu), yunchao@illinois.edu (Y. Wei), 20090964@bit.edu.cn (J. Li), shkwei@bjtu.edu.cn (S. Wei).

https://doi.org/10.1016/j.neucom.2019.06.079
0925-2312/

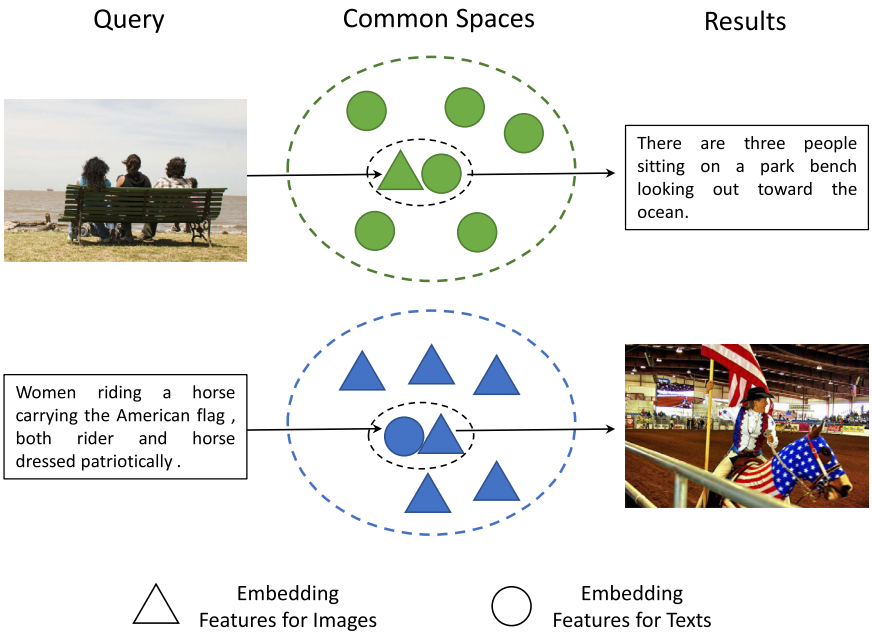


Fig. 1. The motivation of this work. We target to learn task-specific common spaces for different cross-modal retrieval tasks. Triangles and circles indicate embedding features for images and texts, respectively. The green and blue indicate the learned common spaces for I2T and T2I, respectively.

To address the raised issues, we propose a new Adversarial Task-Specific Learning (ATSL) approach to learn distinct embeddings for different retrieval tasks. As shown in Fig. 1, our target is to learn two different common spaces for two retrieval tasks similar to [7]. The two colors (*green* and *blue*) indicate that the common spaces are learned in different manners even the embedded feature representations are with the same dimension. To overcome the drawbacks raised by Wei et al. [7], the following strategies are adopted. First of all, we leverage mid-level attributes instead of high-level categories to encourage semantic consistent of pair-wise embedding. Since attributes can be easily obtained from textual data, our approach has a better generalization ability for both supervised and unsupervised cases compared with the method using human annotations. Secondly, the popular triplet optimization is embedded in our ATSL, so that the pair-wise items are usually with more similarity in terms of distance metrics compared with the unrelated ones. Finally, to guarantee the distribution consistent, we introduce an adversarial learning component to better narrow the distribution differences of the learned embedding features of images and texts. Our ATSL is integrated into a unified deep-based framework, which can be effectively trained in an end-to-end manner.

In summary, main contributions of the paper can be summarized as follows:

- We leverage the semantic attributes to encourage the learned common feature embeddings of couples to be semantic consistent.
- We employ the triplet optimization to guarantee that similar items from different modalities are with smaller distance in the learned common space compared with the dissimilar ones.
- The least squares generative adversarial networks are introduced to learn common feature embeddings for different modalities, in which the inconsistent distribution can be relieved.
- A task-specific learning manner is adopted to produce better optimized common feature embeddings for different retrieval tasks.
- Extensive experiments on the two publicly available datasets (MS COCO and Flicker 30K) significantly demonstrate that the proposed method outperforms other state-of-the-art methods.

2. Related work

2.1. Cross-modal retrieval

Inspired by *CCA* [8], many interesting works [4,9–11] are proposed to improve the performance for multi-modal retrieval. Specifically, Yang et al. [12] propose cross-modal retrieval framework based on semi-supervised learning, which is applicable to many applications including 3D pose/motion data retrieval, image retrieval and cross-modal retrieval. Moreover, Andrew et al. introduce a deep *CCA* (*DCCA*) [13] to learn complex nonlinear embedding for two associated views. Gong et al. [14] first propose to learn three-view embedding for capturing high-level semantic information by supervised and unsupervised manner. Feng et al. [15] employ three kinds of correspondence autoencoders to learn both representation and correlation in a single model. Klein et al. [16] focus on encoding word2vec embedding by exploiting Laplacian mixture model and hybrid Gaussian-Laplacian mixture model. Wang et al. [17] apply two fully connected layers to learning cross-modal structure-preserving embeddings for text and image, respectively. Eisenschtat and Wolf [18] propose a tied 2-way neural network framework for image-sentence matching task. Wang et al. [19] propose adversarial cross-modal retrieval method to learn both discriminative and modality invariant representation for image-sentence matching tasks. Wang et al. [20] propose embedding network and similarity network to learn common latent embedding space and predict a similarity score for cross-modal retrieval tasks. Cao et al. [21] exploit multi-view nonparametric discriminant analysis for class-level cross-modal retrieval. Ren and Hua [22] learn to integrate local concepts with their geometry structures as the side information for image captioning task. Huang and Peng [5] transfer knowledge from a large labeled cross-modal dataset to relatively small dataset by two-level transfer architecture and progressive transfer learning mechanism. Wu et al. [23] propose a novel two-stage approach to learn cross-modal

embeddings by mapping cross-modal data to low dimensional subspace that preserve both semantic and feature information. Huang et al. [24] enhance the image representation by learning semantic attributes and incorporating them into a correct semantic order for cross-modal retrieval task. Liu et al. [25] deal with cross-modal hashing task by fully exploiting the complementary information and semantic correlation, which is used to preserve correlation between categories and heterogeneous instances. Gu et al. [26] incorporate I2T and T2I generative models into the deep conventional cross-modal feature embedding to improve the performance of cross-modal retrieval. Yu et al. [27] propose a regularized cross-modality ranking (*ReCMR*) method to tackle cross-modal zero-shot learning task by taking both inter-modal semantics and intra-modal into consideration. Cao et al. [28] propose a novel multi-view modular discriminant analysis approach to generalize multi-view embedding method using the graph embedding framework for class-level image-sentence matching. Zhang and Lu [29] propose an image-text classification (*CMPC*) loss and image-text matching loss (*CMPM*) to learn discriminative cross-modal representation for image and sentence. Chowdhur et al. [30] propose a two-stage approach for mining weakly-annotate web images to learn a more robust visual-semantic embedding. Liu et al. [31] propose to learn a modality-invariant text-image embedding for image-text matching by involving adversarial learning.

2.2. Generative adversarial nets

Goodfellow et al. [32] first propose a magical framework (*GAN*) for estimating generative models by an adversarial learning process, in which they train two parts: a generator *G* that captures the data distribution from the training dataset, and a discriminator *D* that estimates the probability that a sample is from the real training dataset rather than the generator Mirza and Osindero [33] propose the conditional generative adversarial nets (*CGAN*), which can be built by simply feeding the conditional variable for both the discriminator and generator. Radford et al. [34] introduce a stable framework for training deep generative adversarial networks via *CNN*. Reed et al. [35] investigate *GAN* formulation and a novel deep framework to deal with the text-to-image synthesis task. Isola et al. [36] develop conditional adversarial networks named *pix2pix* for image-to-image translation problems. Ledig et al. [37] present *SRGAN*, an adversarial network (*GAN*) for photo-realistic single image super-resolution (*SR*). Mao et al. [38] propose the least squares generative adversarial networks (*LSGANs*) that utilize the least squares loss function when training the discriminator. Choi et al. [39] propose *StarGAN*, a scalable and novel approach that can perform cross-domain image-to-image translations by exploiting only one model. Arjovsky et al. [40] introduce an algorithm that deems *WGAN*, an alternative to traditional *GAN* training. Zhang et al. [41] develop the self-attention generative adversarial network (*SAGAN*) that allows long-range and attention-driven dependency modeling for image generation tasks. Gulrajani et al. [42] propose an alternative approach to clip weights based on *WGAN*: penalizing the norm of gradient of the critic with respect to its input. Brock et al. [43] train *BigGAN* at the largest scale yet attempted, and analyse of the training behavior of large scale GANs.

2.3. Deep metric learning

Metric learning learns a metric function from training data to calculate the similarity or distance between samples, which is a key component for cross-modal retrieval. Yi et al. [44] present a deep metric learning method (*DML*) by making use of siamese convolutional neural network to learn a similarity metric from image pixels. Hu et al. [45] propose a discriminative deep metric learning approach (*DDML*) for face verification task in the wild. Song et al. [46] introduce a method by taking full use of the training mini-batch by lifting the vector of pairwise distances within the mini-batch to the matrix of pairwise distances. What's more, Hu et al. [47] propose a deep metric learning (*DTML*) method for domain-transfer visual recognition. Sohn [48] focus on a novel metric learning objective named *multi-class N-pair loss*. Liong et al. [49] propose a new deep coupled metric learning (*DCML*) approach for image-text matching, which focus on matching samples captured from two different modalities. Chen et al. [50] consider video object segmentation problem as a pixel-wise search in an embedding subspace via the modified triplet loss function. Duan et al. [51] propose a deep adversarial metric learning (*DAML*) architecture to generate synthetic hard negatives based on the observed negative samples, which can be widely applied to existing supervised deep metric learning approach. Qian et al. [52] adopt the margin preserving metric learning framework (*MaPML*) to learn both similarity metric and latent samples. Iscen et al. [53] present a novel unsupervised framework for hard training example mining on Manifolds. Zhao et al. [54] propose a hard triplet generation method (*HTG*) via adversarial training for learning an optimal feature embedding for images. Xie et al. [55] attempt to address three issues of existing orthogonality-promoting *DML* approaches that include lacking theoretical analysis and computational inefficiency in generalization.

3. Adversarial task-specific learning

In this section, we give an overview of the proposed adversarial task-specific learning approach, and each component in details will be illustrated then.

3.1. Overview

The goal of ATSL is to find a discriminative common space where the distance of data from different modalities can be measured effectively. Concretely, we first extract their feature vectors for the visual and text data, respectively. Then, features in heterogeneous space are simultaneously fed into the proposed network to learn the optimal feature representations. The overview of the proposed ATSL is illustrated in Fig. 2, which consists of three novel components, i.e. neighborhood-preserving image-text embedding, modality-dependent attribute-preserving feature learning and adversarial cross-modal feature learning. Since the ATSL is based on the generative adversarial networks, the training stage is divided into two parts (generator and discriminator), and the discriminator is trained before the generator. During the process of training generator, we jointly minimize the embedding, adversarial and classification losses. Thus, the loss function for the discriminator can be formulated as:

$$\mathbb{L}_D = L_{lsgan}(D) \tag{1}$$

where $L_{lsgan}(D)$ is the least squares loss for the discriminator.

For the generator, the loss can be defined as:

$$\mathbb{L}_G = L_{emd} + \lambda \cdot L_{attr} + \mu \cdot L_{lsgan}(G) \tag{2}$$

where L_{emd} and L_{attr} are the loss functions for embedding and attributes classification respectively. the $L_{lsgan}(G)$ is the least squares loss for the generator. λ and μ decide the weights of the attributes loss and the least squares loss.

3.2. Triplet-based neighborhood preserving learning

In cross-modal retrieval task including sentence-to-image and image-to-sentence search, it is of significance to learn a

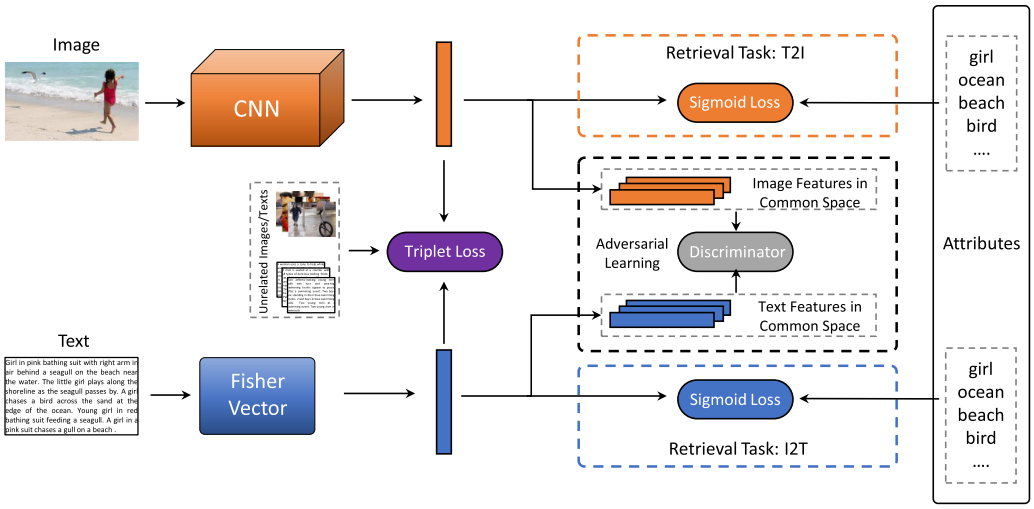


Fig. 2. Overview of the adversarial task-specific learning. We first extract features of texts and images, and map them into common subspace. Triplet loss is then exploited in order to make the paired image/text as close as possible and the unpaired image/text as far as possible. We utilize generative adversarial network to keep the distribution of text and image consistent. Texts' attributes are preserved for image-to-text retrieval tasks, but we retain the attributes of images if texts retrieve images.

heterogeneous semantic representation for text and image. To be specific, given an sentence (image), the goal is to find the best matching image (sentence) from the database. The task can be handled identically by nearest neighbor search in a common image-sentence embedding space. Consequently, we take triplet-based neighborhood preserving learning into consideration to minimize the gap between pairwise cross-modal datas while maximizing the distance between semantically different datas. We enforce triplet constraints into the optimizing function for learning to rank via a triplet loss term.

Given a training image I_x, T_y and T_z indicate the sets of pairwise matching (positive) and non-matching (negative) texts, respectively. Analogously, given a sentence $T_{x'}$, $I_{y'}$ and $I_{z'}$ indicate the sets of pairwise matching and non-matching images. $T_{y''}$ describes the similar image with $T_{x''}$. Note that $T_{z''}$ doesn't share the similar image with $T_{x''}$. Three optimization terms are employed for embedding the isomorphic semantic representations. In general, the objective function used for optimizing is defined as follow:

$$L_{emd} = \alpha \cdot L_{img} + \beta \cdot L_{sent} + \gamma \cdot L_{sent_{neighbor}} \tag{3}$$

We take the image and text as anchor to get the following two sets of triplet loss L_{img} and L_{sent}:

$$L_{img} = \sum_{I_x, T_z} max\Big(mg + dist\Big(f^{(I)}_{W_I}(I_x), f^{(T)}_{W_T}(T_y)\Big) - dist\Big(f^{(I)}_{W_I}(I_x), f^{(T)}_{W_T}(T_z)\Big)\Big) \tag{4}$$

$$L_{sent} = \sum_{T_{x'}, I_{z'}} max\Big(mg + dist\Big(f^{(T)}_{W_T}(T_{x'}), f^{(I)}_{W_I}(I_{y'})\Big) - dist\Big(f^{(T)}_{W_T}(T_{x'}), f^{(I)}_{W_I}(I_{z'})\Big)\Big) \tag{5}$$

Neighborhood preserving embedding between texts can be formulated as:

$$L_{sent_{neighbor}} = \sum_{T_{x''}, T_{z''}} max\Big(mg + dist\Big(f^{(T)}_{W_T}(T_{x''}), f^{(T)}_{W_T}(T_{y''})\Big) - dist\Big(f^{(T)}_{W_T}(I_{x''}), f^{(T)}_{W_T}(T_{z''})\Big)\Big) \tag{6}$$

where $dist(X, Y)$ indicates the *Euclidean* distance between X and Y in the embedding space. $f^{(T)}_{W_T}$ denotes the embedding text feature in the common space, and $f^{(I)}_{W_I}$ denotes the embedding image feature. The margin of the triplet is mg. The weights α and β and γ balance the strength of the ranking loss in each part.

3.3. Modality-dependent attribute-preserving feature learning

Since the text or image is mapped into a common space, it is unavoidable to lose some information. We propose to maintain such semantic attribute abstraction in the shared subspace for each modality. In such a case, the embedded features from each domain is encouraged to predict categories accurately simultaneously. Specifically, we reach this purpose through the multi-label semantic attributes classification.

We regard the semantic categories prediction as a multi-label classification task by minimizing the sigmoid cross-entropy loss function as:

$$L_{attr} = -\frac{1}{n}\sum_{i=0}^{n}[p_i log\hat{p}_i + (1 - p_i)log(1 - \hat{p}_i)] \tag{7}$$

where $\hat{p}_i$ and p_i are the prediction of the network and the target of label i, respectively. For image-to-sentence retrieval task, we make images's attributes invariant when being embedded to common space. Attributes of sentences in common space remain constant from sentence-to-image retrieval.

3.4. Adversarial cross-modal feature learning

It is a common knowledge that the features of images and sentences have inconsistent distribution and representation because they come from different domain. Accordingly, bridging the domain gap in a common space for images and sentences is of great significance. To achieve this goal, we exploit generative adversarial networks (GANs) to keep their distribution consistent in the common space. In our framework, the least squares generative adversarial networks (LSGANs) [38] are adopted as the backbone of the framework. LSGANs can move the generated samples toward the decision boundary by penalizing samples locating on the right side of the decision boundary, which is the main benefit.

GANs are to train a discriminator D and a generator G interchangeably, which aim at making the distribution of text (T) same as that of image (I). Here, G is feature embedding (f_W) in *section 3.2*. Consequently, the objective loss functions for adversarial feature learning can be defined as follows:

Table 1
The parameters of our model for different tasks.

dataset	modality	μ	λ	lrD	lrG	Batchsize	negsample	mg	α	γ	β
Flickr30K	image modality	0.01	0.02	0.0001	0.0004	500	10	0.02	1.5	0.02	1.6
	sentence modality	0.01	0.06	0.0001	0.0003	500	10	0.05	1.5	0.05	1
MS COCO	image modality	0.01	0.02	0.0001	0.0004	500	10	0.02	1.5	0.02	1.6
	sentence modality	0.01	0.06	0.0001	0.0003	500	10	0.05	1.5	0.05	1

$$\min_D L_{lsgan}(D) = \frac{1}{2}\mathbb{E}_{I\sim p_{data}(I)}[(D(f_{W_I}^{(I)}(I)) - 1)^2] + \frac{1}{2}\mathbb{E}_{T\sim p_{data}(T)}[(D(f_{W_T}^{(T)}(T)))^2] \quad (8)$$

$$\min_G L_{lsgan}(G) = \frac{1}{2}\mathbb{E}_{T\sim p_{data}(T)}[(D(f_{W_T}^{(T)}(T)) - 1)^2] \quad (9)$$

For image-to-sentence retrieval task, image domain is the target domain, while sentence domain is the target domain at the sentence-to-image retrieval task.

The optimization of our framework is shown in the Algorithm 1.

Algorithm 1: Pseudocode of our ATSL.

Initialization:
Image features in minibatch: $\mathbb{I} = \{I_1, I_2, \ldots, I_n\}$;
Text features in minibatch: $\mathbb{T} = \{T_1, T_2, \ldots, T_n\}$;
Attribute Label: $P = \{p_1, p_2, \ldots, p_n\}$;
Hyperparameters: $\alpha, \beta, \gamma, \lambda, \mu, mg, negsample, lrD, lrG$;
for *number of training iterations* **do**
  for *k steps* **do**
    Update the parameters of discriminator by descending its stochastic gradient with learning rate *lrD*:
    $\nabla_W(L_{lsgan}(D))$
  end
  Update the parameters of embedding and attribute learning by descending its stochastic gradient with learning rate *lrG*:
  $\nabla_W(L_{emd} + \lambda \cdot L_{attr} + \mu \cdot L_{lsgan}(G))$
end
return learned embedding feature in common subspace: $f_{W_I}^{(I)}$ and $f_{W_T}^{(T)}$;

4. Experiment results

4.1. Dataset and evaluation metrics

To evaluate the effectiveness of proposed approach, we conduct extensive experiments on two popular available datasets, *i.e.* Flickr30K and MS COCO.

Flickr30K: This dataset consists of 30,783 images, and each image is accompanied by five descriptive sentences. Following the same protocol as the recent work [17,20], we randomly split it into a training set and a test set with 29,783 and 1000 couples, respectively.

MS COCO: We use the MS COCO caption dataset which consists of 114,287 images and accompanies by five descriptive sentences for each image as well. We adopt the same splits as reported in [17,20], 1000 couples used for testing and the rest for training.

Evaluation metrics: In our experiments, T2I and I2T tasks are both considered. We report the performance at Recall@K (K = 1, 5, 10), which is the percentage of queries that at least one correct result is ranked among the top K of the ranked list.

4.2. Training/testing settings

To encode images, we exploit VGG19 convolutional neural network model to obtain 4096-dimensional feature, which is similar to [17]. Firstly, we resize the images to 256 pixels × 256 pixels. Then, we crop each image into ten different 224 pixels × 224 pixels images by mining the center, the four corners and them on *x*-axis mirror symmetry. Next, we subtract the average strength from each color channel of RGB image. Finally, the image is encoded by the mean outputs of the resulting images. As for sentences representation, we extract 18000-dimensional fisher vector representation by exploiting hybrid gaussian-laplacian mixture model, which is proposed by [16]. To save memory and training time, we reduce these 18000-dimensional vectors to 6000 dimensions by multiplying random matrices.

For both Flickr30K and MS COCO datasets, we apply 4096 → 2048 → 512 embedding on the image side when exploiting 4096-dimensional visual features. On the text side, the output dimensions of two fully connected layers are [2048, 512]. For attribute-preserving branch, we also utilize two FC layers whose dimensions are [512, 512] and stick to the two FC layers (512 → 256 → 1) for the adversarial learning.

The most frequent 512 words are selected to form a vocabulary after eliminating stem words and stop words. According to this dictionary, we represent a given text as a vector with multiple labels. For a given text, it is set to 0 if a word does not exist in the dictionary, otherwise it is 0. If no word in this sentence exists in the dictionary, all words are set to -1.

In this paper, all attribute vectors are constructed automatically without any manual annotation. The most frequent 512 words are selected to form a vocabulary after eliminating stem words and stop words. According to this dictionary, we represent a given text as a vector with multiple labels. For a given text, it is set to 0 if a word does not exist in the dictionary, and vice versa, otherwise it is 0. If no word in this sentence exists in the dictionary, all words are set to -1.

We train different retrieval tasks with different initialization parameters. The parameters of different tasks for different datasets are summarized in Table 1. In this paper, experiments for two modalities, i.e., image modality and sentence modality, are conducted on Flickr30K and MS COCO. For the same modality, the hyper-parameters for both datasets are same. We determine the hyper-parameters through a series of experiments carried out on Flickr30K due to its samller size. From Table 2, it is obvious that the second row achieves the best performance of sentence-image retrieval, where λ, *lrD*, *mg*, β and γ are set to 0.02, 0.0004, 0.02, 1.6 and 0.02, respectively. The lowest SUM score lies in the 15th row, where all parameters are consistent with the second row except γ. Therefore, γ is considered as a factor that has the greatest impact on overall performance. For sentence modality, we can obtain the optimum performance by setting the parameters to 0.06, 0.0003, 0.05, 1 and 0.05 as shown in the third row of Table 3. In contrast, margin of triplet loss is the most critical factor for performance of image-sentence retrieval for sentence modality, which is demonstrated in the ninth row of Table 3. Our networks are trained and tested on NVIDIA GeForce TITAN X GPU with

Table 2
Retrieval results of different hyper-parameters for image modality on Flickr30K.

Flickr30K/image modality	sentence-image				
	R@1	R@5	R@10	R@50	SUM
1.$\lambda = 0.01, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.353	0.668	0.763	**0.963**	2.747
2.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.02$	**0.386**	0.673	0.778	0.953	**2.790**
3.$\lambda = 0.06, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.363	0.671	0.776	0.955	2.765
4.$\lambda = 0.08, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.385	0.672	0.774	0.952	2.783
5.$\lambda = 0.02, lrD = 0.0002, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.368	0.672	0.773	0.960	2.773
6.$\lambda = 0.02, lrD = 0.0003, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.375	0.669	0.777	0.953	2.774
7.$\lambda = 0.02, lrD = 0.0005, mg = 0.02, \beta = 1.6, \gamma = 0.02$	0.367	0.674	0.778	0.952	2.771
8.$\lambda = 0.02, lrD = 0.0004, mg = 0.01, \beta = 1.6, \gamma = 0.02$	0.372	0.672	0.770	0.948	2.762
9.$\lambda = 0.02, lrD = 0.0004, mg = 0.05, \beta = 1.6, \gamma = 0.02$	0.367	**0.677**	**0.780**	0.955	2.779
10.$\lambda = 0.02, lrD = 0.0004, mg = 0.08, \beta = 1.6, \gamma = 0.02$	0.367	0.670	0.773	0.951	2.761
11.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 0.6, \gamma = 0.02$	0.3 80	0.663	0.776	0.953	2.772
12.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 1.0, \gamma = 0.02$	0.382	0.667	0.778	0.948	2.775
13.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 2.0, \gamma = 0.02$	0.362	0.673	0.774	0.958	2.767
14.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.01$	0.365	0.666	0.773	0.958	2.762
15.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.06$	0.373	0.663	0.767	0.953	2.756
16.$\lambda = 0.02, lrD = 0.0004, mg = 0.02, \beta = 1.6, \gamma = 0.08$	0.367	0.667	0.777	0.958	2.769

Table 3
Retrieval results of different hyper-parameters for sentence modality on Flickr30K.

Flickr30K/sentence modality	image-sentence				
	R@1	R@5	R@10	R@50	SUM
1.$\lambda = 0.01, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.05$	0.360	0.654	0.774	0.948	2.736
2.$\lambda = 0.02, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.05$	0.359	0.653	0.773	0.946	2.731
3.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.05$	**0.384**	0.672	0.780	**0.948**	**2.784**
4.$\lambda = 0.08, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.05$	0.354	0.663	**0.781**	0.945	2.743
5.$\lambda = 0.06, lrD = 0.0002, mg = 0.05, \beta = 1, \gamma = 0.05$	0.354	0.656	0.774	0.940	2.724
6.$\lambda = 0.06, lrD = 0.0004, mg = 0.05, \beta = 1, \gamma = 0.05$	0.355	0.667	0.772	0.938	2.732
7.$\lambda = 0.06, lrD = 0.0005, mg = 0.05, \beta = 1, \gamma = 0.05$	0.356	0.660	0.777	0.940	2.733
8.$\lambda = 0.06, lrD = 0.0003, mg = 0.01, \beta = 1, \gamma = 0.05$	0.371	0.661	0.776	0.939	2.747
9.$\lambda = 0.06, lrD = 0.0003, mg = 0.02, \beta = 1, \gamma = 0.05$	0.354	0.645	0.769	0.943	2.711
10.$\lambda = 0.06, lrD = 0.0003, mg = 0.08, \beta = 1, \gamma = 0.05$	0.374	0.676	0.774	0.946	2.770
11.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 0.6, \gamma = 0.05$	0.364	0.664	0.780	0.935	2.743
12.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 1.6, \gamma = 0.05$	0.362	0.665	0.771	0.942	2.740
13.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 2, \gamma = 0.05$	0.366	0.679	0.783	0.943	2.771
14.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.01$	0.360	0.660	0.777	0.935	2.732
15.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.02$	0.364	**0.681**	0.773	0.947	2.765
16.$\lambda = 0.06, lrD = 0.0003, mg = 0.05, \beta = 1, \gamma = 0.07$	0.367	0.666	0.774	0.938	2.745

12GB memory, which is based on the publicly available tensorflow framework [56].

4.3. Comparisons with state-of-the-arts

We compare our ATSL approach with state-of-the-art methods on MS COCO and Flickr30K datasets, which have been widely adopted as benchmark datasets in the latest papers. In order to make a fair comparison with [16,17,20], we report the results by running released code under the same feature representation, which is introduced at *Training/Testing Settings*.

We compare the number of parameters with Two-Branch [20], which yields the best perfomance among the references. In the testing phase, like Two-Branch [20], we apply 4096 → 2048 → 512 embedding on the image side and 6000 → 2048 → 512 embedding on the text side, which leads to a same computational cost as Two-Branch [20]. The number of parameters is only 0.288% more than Two-Branch for training.

Table 4 shows the comparisons on MS COCO. It can be observed that our ATSL achieves the state-of-the-art results in terms of all the evaluation metrics. In particular, ATSL[sent] and ATSL[img] achieve 0.571 and 0.565 in R@1 for I2T and T2I tasks, which outperform other approaches more than 3.2% and 3.0%, respectively. Two-Branch method is the most similar work to ours, which utilizes two network structures to embed representations for different modalities and takes neighborhood information into consideration. Besides, Structure Preserving approach proposes a method that learning joint embeddings by introducing neighborhood structure preservation constraint with a two-branch neural network, which has a little parallelism with our methods. However, all the mentioned methods do not take the distribution consistency of different modalities into consideration. In addition, we apply a task-specific learning manner to obtain different common spaces for different tasks, which can achieve the best performance for both T2I and I2T.

Table 5 shows the comparisons on Flick30K. We can see that our ATSL outperforms prior methods by a relatively large margin. In particular, ATSL[sent] and ATSL[img] lead to 0.384 and 0.386 in R@1 for I2T and T2I tasks, which obtain significant improvements over other approaches more than 4.3% and 3.6%, respectively. In addition, our method can also gain more than 2% at Recall@5.

Both experimental results on MS COCO and Flickr30K well demonstrate the superiority of our approach. Finally, we visualize some T2I and I2T examples in Figs. 3 and 4, respectively. It can be observed that our approach obtains the most ideal results.

4.4. Ablation study

To validate the effectiveness of each component in our proposed ATSL, we conduct ablation studies on Flickr30K and MS COCO datasets.

Table 4
Bidirectional retrieval results on MS COCO 1000-image test set.

MS COCO	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
ACMR[ACM2017] [19]	–	–	–	0.932	–	–	–	0.871
Corr-AEs[ACM2014] [15]	0.264	0.609	0.759	0.97	0.26	0.603	0.675	0.961
Cross-corr-AEs[ACM2014] [15]	0.289	0.615	0.756	0.964	0.286	0.599	0.751	0.956
Full-corr-AEs[ACM2014] [15]	0.29	0.629	0.772	0.975	0.281	0.637	0.767	0.964
CCA[FV HGLMM CVPR2015] [16]	0.311	0.616	0.755	0.969	0.253	0.535	0.675	0.961
Structure Preserving[CVPR2016] [17]	0.487	0.813	0.894	0.988	0.476	0.820	0.895	0.983
VSEVGG19[ACM2018] [30]	0.537	–	0.925	–	0.412	–	0.897	–
Two-Branch[PAMI2018] [20]	0.539	0.844	0.913	0.99	0.535	0.843	0.919	0.992
ATSL[sent]	**0.571**	**0.848**	**0.932**	**0.991**	–	–	–	–
ATSL[img]	–	–	–	–	**0.565**	**0.855**	**0.933**	**0.995**

Table 5
Bidirectional retrieval results on Flickr30K 1000-image test set.

Flickr30K	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Corr-AEs[ACM2014] [15]	0.234	0.521	0.647	0.877	0.222	0.49	0.608	0.847
Cross-corr-AEs[ACM2014] [15]	0.221	0.503	0.615	0.88	0.219	0.469	0.61	0.872
Full-corr-AEs[ACM2014] [15]	0.267	0.551	0.673	0.899	0.248	0.527	0.653	0.874
CCA[FV HGLMM CVPR2015] [16]	0.329	0.607	0.736	0.929	0.304	0.625	0.734	0.927
Structure Preserving[CVPR2016] [17]	0.329	0.609	0.727	0.911	0.328	0.637	0.728	0.913
VSEVGG19[ACM2018] [30]	0.378	–	0.771	–	0.279	–	0.689	–
Two-Branch[PAMI2018] [20]	0.341	0.653	0.753	0.941	0.35	0.646	0.763	0.943
ATSL[sent]	**0.384**	**0.672**	**0.78**	**0.948**	–	–	–	–
ATSL[img]	–	–	–	–	**0.386**	**0.673**	**0.778**	**0.953**

Fig. 3. Examples of comparisons with state-of-the-arts for text-to-image retrieval. For each text query, the top-1 ranked image is shown.

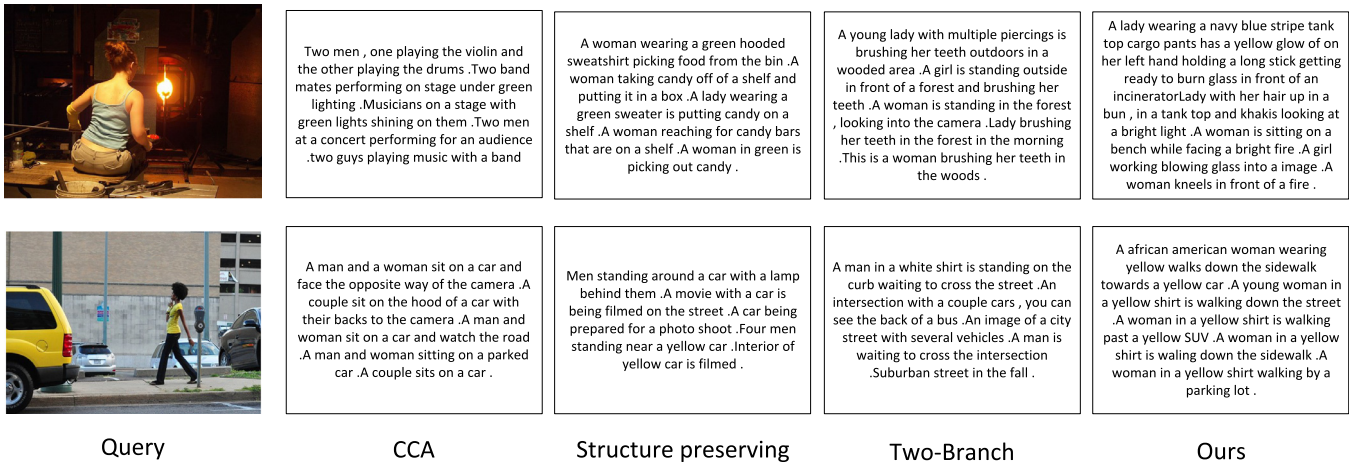


Fig. 4. Examples of comparisons with state-of-the-arts for image-to-text retrieval. For each image query, the top-1 ranked text is shown.

Table 6
Attribute-preserving retrieval results on MS COCO 1000-image test set.

MS COCO	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Two-Branch[PAMI2018] [20]	0.539	0.844	0.913	0.99	0.535	0.843	0.919	0.992
Two-Branch+attr[sent]	0.548	0.853	0.932	0.992	0.547	0.855	0.926	0.992
Two-Branch+attr[img]	0.543	0.845	0.928	0.992	0.559	0.857	0.924	0.992
ATSL[sent]	**0.571**	0.848	**0.932**	0.991	0.555	**0.857**	0.93	0.993
ATSL[img]	0.559	**0.852**	0.929	**0.993**	**0.565**	0.855	**0.933**	**0.995**

Table 7
Attribute-preserving retrieval results on Flickr30K 1000-image test set.

Flickr30K	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Two-Branch[PAMI2018] [20]	0.341	0.653	0.753	0.941	0.35	0.646	0.763	0.943
Two-Branch+attr[sent]	0.356	0.656	0.769	0.941	0.363	0.647	0.761	0.947
Two-Branch+attr[img]	0.357	0.663	0.77	0.941	0.357	0.654	0.762	0.951
ATSL[sent]	**0.384**	0.672	0.78	**0.948**	0.376	0.667	**0.78**	**0.953**
ATSL[img]	0.374	**0.679**	**0.781**	0.941	**0.386**	**0.673**	0.778	**0.953**

Table 8
Adversarial learning results on MS COCO 1000-image test set.

MS COCO	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Two-Branch[PAMI2018] [20]	0.539	0.844	0.913	0.99	0.535	0.843	0.919	0.992
Two-Branch+GAN[sent]	0.552	0.854	0.924	0.99	0.547	0.855	0.919	0.992
Two-Branch+GAN[img]	0.554	0.85	0.922	0.993	0.559	0.848	0.921	0.992
ATSL[sent]	**0.571**	0.848	**0.932**	0.991	0.555	**0.857**	0.93	0.993
ATSL[img]	0.559	**0.852**	0.929	**0.993**	**0.565**	0.855	**0.933**	**0.995**

Table 9
Adversarial learning results on Flickr30K 1000-image test set.

Flickr30K	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Two-Branch[PAMI2018] [20]	0.341	0.653	0.753	0.941	0.35	0.646	0.763	0.943
Two-Branch+GAN[sent]	0.37	0.653	0.769	0.942	0.353	0.657	0.767	0.947
Two-Branch+GAN[img]	0.365	0.662	0.77	0.942	0.38	0.676	0.767	0.953
ATSL[sent]	**0.384**	0.672	0.78	**0.948**	0.376	0.667	**0.78**	**0.953**
ATSL[img]	0.374	**0.679**	**0.781**	0.941	**0.386**	**0.673**	0.778	**0.953**

4.4.1. Attribute-preserving feature learning

The introduced attribute-preserving feature learning encourages the learned feature embeddings to be semantic consistent. To analyze this component, we adopt neighborhood preserving embedding as the baseline model [20]. Hence, we conduct experiments to validate the effectiveness of attribute-preserving feature learning on the two datasets.

Table 6 shows the results validated on MS COCO, it is obvious that our proposed method outperforms the baseline model significantly, especially at Recall@1. The results on Flicker 30K are shown in Table 7, it is better to learn attribute-preserving embedding than not. With attribute-preserving, the T2I and I2T yield results of 0.357/0.363 in Recall@1, respectively. All the experiment results demonstrate the effectiveness of the attribute-preserving feature learning for cross-model retrieval task. In this paper, we want to highlight that we can obtain the optimal performance for both text-to-image retrieval and image-to-text retrieval by training the networks in different ways, which can motivate other researchers to balance the two tasks. In addition, for single task, we have the same computational efficiency as the Two-branch method in the testing phase. It is more concerned about the retrieval time of offline than the training time of online for cross-modal retrieval task. Moreover, 3% performance improvement is relatively high for shallow features, which is shown in Tables 4 and 5.

4.4.2. Adversarial cross-modal feature learning

In this section, we analyze the adversarial cross-modal feature learning. During the period of mapping the cross modal data into a common subspace, the data distribution is usually not taken into account, which would cause semantic deviation adverse to the retrieval task. Therefore, we employ the generative adversarial networks to keep the two distributions consistent. To prove the effectiveness of adversarial learning, we conduct experiments with and without training LSGANs.

From the results on MS COCO dataset shown in Table 8, it can be noted that our method improves on all the evaluation metrics and makes a breakthrough at Recall@1. In particular, the approach with adversarial learning yields the Recall@1 accuracy by approximately 2.4% for sentence-to-image retrieval and 1.5% for image-to-sentence retrieval. These experimental results further validate the effectiveness of the proposed adversarial feature learning method for cross-model retrieval.

Table 9 reports the effectiveness of adversarial learning on Flickr30K. From Table 9, we can see that the Recall@1 score is improved from 0.341 to 0.37 for I2T and 0.35 to 0.353 for T2I with the sentence attribute preserved. With the adversarial learning, Recall@1 score is increased from 0.35 to 0.38, with 3% improvement for sentence-to-image retrieval, and the performance of image-to-sentence retrieval is improved as well.

Table 10
Task-specific retrieval results on Flickr30K 1000-image test set.

Flickr30K	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
ATSL[double attributes][sent]	0.371	0.669	0.78	0.946	0.381	0.665	0.761	0.951
ATSL[double attributes][img]	0.37	0.661	0.766	0.936	0.371	0.656	0.765	0.942
ATSL[sent]	**0.384**	0.672	0.78	**0.948**	0.376	0.667	**0.78**	**0.953**
ATSL[img]	0.374	**0.679**	**0.781**	0.941	**0.386**	**0.673**	0.778	**0.953**

Table 11
Task-specific retrieval results on MS COCO 1000-image test set.

MS COCO	image-sentence				sentence-image			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
ATSL[double attributes][sent]	0.553	0.85	0.928	0.99	0.549	0.852	0.934	0.989
ATSL[double attributes][img]	0.558	0.845	0.926	0.993	0.559	0.847	0.927	0.993
ATSL[sent]	**0.571**	0.848	**0.932**	0.991	0.555	**0.857**	0.93	0.993
ATSL[img]	0.559	**0.852**	0.929	**0.993**	**0.565**	0.855	**0.933**	**0.995**

4.4.3. Task-specific learning

Tables 10 and 11 evaluate the effectiveness of task-specific learning on Flickr30K and MS COCO, respectively. ATSL with double attributes is trained with attributes both on images and texts. As shown in two tables, we can see that the scores of ATSL are higher than that of ATSL with double attributes on both image-to-sentence and sentence-to-image retrieval. Consequently, our ATSL algorithm is more suitable to deal with image-sentence retrieval tasks.

5. Conclusion

In this paper, we propose an adversarial task-specific learning approach for image-text retrieval, which can not only make the distribution of image and text consistent but also preserve the attribute in common subspace. Specially, this approach can learn distinct embeddings for different retrieval tasks, i.e. images retrieve texts (I2T) or texts retrieve images (T2I). Our framework is divided into three parts. Firstly, we leverage mid-level attributes instead of high-level categories to encourage semantic consistency of pair-wise embedding. Then, the popular triplet optimization is embedded in our ATSL, so that the pair-wise items are usually more similar in terms of distance metrics compared with the unrelated ones. Finally, we introduce an adversarial learning component to better narrow the distribution differences of the learned embedding features of images and texts. Our proposed framework yields state-of-the-art results on MS COCO dataset and Flickr30K dataset. Future research directions include integrating word-level, phrase-level and sentence-level matching to learn rich feature embeddings.

Disclosure of conflicts of interest

We confirm that there are no conflicts of interest associated with this publication. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from yzhao@bjtu.edu.cn

Acknowledgment

The authors must thank the anonymous reviewers for their constructive comments and valuable suggestions on this paper. This work was jointly supported by the National Key Research and Development of China (No.2016YFB0800404), the National Science Foundation of China (No.61532005, No.61572065), Program of China Scholarships Council (No.201807095006) and the Fundamental Research Funds for the Central Universities (No.2018JBZ001, No.2018YJS028).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neucom.2019.06.079.

References

[1] F. Feng, R. Li, X. Wang, Deep correspondence restricted Boltzmann machine for cross-modal retrieval, Neurocomputing 154 (2015) 50–60.
[2] 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 TCYB 47 (2) (2017) 449–460.
[3] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y.-D. Shen, Dual-path convolutional image-text embedding, 2017. arXiv: 1711.05535.
[4] X. Huang, Y. Peng, M. Yuan, Cross-modal common representation learning by hybrid transfer network, in: Proceedings of the IJCAI, 2017, pp. 19–25.
[5] X. Huang, Y. Peng, Deep cross-media knowledge transfer, in: Proceedings of the IEEE CVPR, 2018, pp. 8837–8846.
[6] M. Xu, Z. Zhu, Y. Zhao, F. Sun, Subspace learning by kernel dependence maximization for cross-modal retrieval, Neurocomputing 309 (2019) 94–105.
[7] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, S. Yan, Modality-dependent cross-media retrieval, ACM TIST 7 (4) (2016) 57.
[8] D.R. Hardoon, S. Szedmak, J. Shawetaylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[9] Y. Peng, J. Qi, CM-GANs: cross-modal generative adversarial networks for common representation learning, ACM TOMM 15 (1) (2018) 22.
[10] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, Y. Zhuang, Deep compositional cross-modal learning to rank via local-global alignment, in: Proceedings of the ACM MM, 2015, pp. 69–78.
[11] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Cross-modal retrieval using multi-ordered discriminative structured subspace learning, IEEE TMM 19 (6) (2017) 1220–1233.
[12] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE TPAMI 34 (4) (2012) 723–742.
[13] G. Andrew, R. Arora, J.A. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the ICML, 2013, pp. 1247–1255.

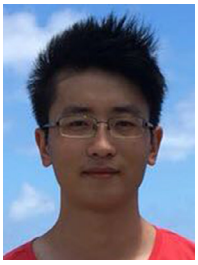
[14] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, IJCV 106 (2) (2014) 210–233.
[15] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the ACM MM, 2014, pp. 7–16.
[16] B. Klein, G. Lev, G. Sadeh, L. Wolf, Associating neural word embeddings with deep image representations using fisher vectors, in: Proceedings of the IEEE CVPR, 2015, pp. 4437–4446.
[17] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: Proceedings of the IEEE CVPR, 2016, pp. 5005–5013.
[18] A. Eisenschtat, L. Wolf, Linking image and text with 2-way nets, in: Proceedings of the IEEE CVPR, 2017, pp. 4601–4611.
[19] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the ACM MM, 2017, pp. 154–162.
[20] L. Wang, Y. Li, J. Huang, S. Lazebnik, Learning two-branch neural networks for image-text matching tasks, IEEE TPAMI 41 (2) (2018) 394–407.
[21] G. Cao, A. Iosifidis, M. Gabbouj, Multi-view nonparametric discriminant analysis for image retrieval and recognition, IEEE SPL 24 (10) (2017) 1537–1541.
[22] L. Ren, K. Hua, Improved image description via embedded object structure graph and semantic feature matching, in: Proceedings of the ISM, 2018, pp. 73–80.
[23] Y. Wu, S. Wang, Q. Huang, Multi-modal semantic autoencoder for cross-modal retrieval, Neurocomputing 331 (2019) 165–175.
[24] Y. Huang, Q. Wu, L. Wang, Learning semantic concepts and order for image and sentence matching, in: Proceedings of the IEEE CVPR, 2018, pp. 6163–6171.
[25] R. Liu, S. Wei, Y. Zhao, Z. Zhu, J. Wang, Multi-view cross-media hashing with semantic consistency, IEEE Multimed. 25 (2) (2018) 71–86.
[26] J. Gu, J. Cai, S. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE CVPR, 2018, pp. 7181–7189.
[27] Y. Yu, Z. Ji, J. Guo, Y. Pang, Zero-shot learning with regularized cross-modality ranking, Neurocomputing 259 (2017) 14–20.
[28] G. Cao, A. Iosifidis, K. Chen, M. Gabbouj, Generalized multi-view embedding for visual recognition and cross-modal retrieval, IEEE TCYB 48 (9) (2018) 2542–2555.
[29] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the ECCV, 2018, pp. 686–701.
[30] M. Chowdhury, P. Rameswar, E. Papalexakis, A. Roy-Chowdhury, Webly supervised joint embedding for cross-modal image-text retrieval, in: Proceedings of the ACM MM, 2018.
[31] R. Liu, Y. Zhao, S. Wei, L. Zheng, Y. Yang, Modality-invariant image-text embedding for image-sentence matching, ACM TOMM 15 (1) (2019) 27.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the NIPS, 2014, pp. 2672–2680.
[33] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014. arXiv: 1411.1784
[34] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: Proceedings of the ICLR, 2016.
[35] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: Proceedings of the ICML, 2016.
[36] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE CVPR, 2017, pp. 1125–1134.
[37] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE CVPR, 2017, pp. 4681–4690.
[38] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE ICCV, 2017, pp. 2813–2821.
[39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE CVPR, 2018.
[40] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: Proceedings of the ICML, 2017, pp. 214–223.
[41] Z. Han, G. Ian, Metaxas dimitris and odena augustus, self-attention generative adversarial networks, in: Proceedings of the ICML, 2019, pp. 7354–7363.
[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of Wasserstein GANs, in: Proceedings of the NIPS, 2017, pp. 5767–5777.
[43] B. Andrew, D. Jeff, S. Karen, Large scale gan training for high fidelity natural image synthesis, in: Proceedings of the ICLR, 2019.
[44] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Proceedings of the IEEE ICPR, 2014, pp. 34–39.
[45] J. Hu, J. Lu, Y.-P. Tan, Discriminative deep metric learning for face verification in the wild, in: Proceedings of the IEEE CVPR, 2014, pp. 1875–1882.
[46] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE CVPR, 2016, pp. 4004–4012.
[47] J. Hu, J. Lu, Y.-P. Tan, Deep transfer metric learning, in: Proceedings of the IEEE CVPR, 2015, pp. 325–333.
[48] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Proceedings of the NIPS, 2016, pp. 1857–1865.
[49] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Deep coupled metric learning for cross–modal matching, IEEE TMM 19 (6) (2017) 1234–1244.
[50] Y. Chen, J. Pont-Tuset, A. Montes, L. Van Gool, Blazingly fast video object segmentation with pixel-wise metric learning, in: Proceedings of the IEEE CVPR, 2018, pp. 1189–1198.
[51] Y. Duan, W. Zheng, X. Lin, J. Lu, J. Zhou, Deep adversarial metric learning, in: Proceedings of the IEEE CVPR, 2018, pp. 2780–2789.
[52] Q. Qian, J. Tang, H. Li, S. Zhu, R. Jin, Large-scale distance metric learning with uncertainty, in: Proceedings of the IEEE CVPR, 2018, pp. 8542–8550.
[53] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Mining on manifolds: metric learning without labels, in: Proceedings of the IEEE CVPR, 2018.
[54] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, X.-s. Hua, An adversarial approach to hard triplet generation, in: Proceedings of the ECCV, 2018, pp. 501–517.
[55] P. Xie, W. Wu, Y. Zhu, E.P. Xing, Orthogonality-promoting distance metric learning: convex relaxation and theoretical analysis, in: Proceedings of the ICML, 2018, pp. 5403–5412.
[56] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, in: Proceedings of the OSDI, 2016, pp. 265–283.

Xin Fu is a Ph.D. candidate of the Institute of Information Science, Beijing Jiaotong University. Her research interests include cross-media retrieval/completion, computer vision and deep learning.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. In October 2015, he visited the Swiss Federal Institute of Technology, Lausanne, Switzerland (EPFL). From December 2017 to March 2018, he visited the University of Southern California. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, video analysis and understanding and artificial intelligence. Dr. Zhao serves on the Editorial Boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, an Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor of Signal Processing: Image Communication. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a Fellow of the IET.

Ting Liu was born in Shannxi, China, in 1993. He received the B.S. degrees in computer science and technology in 2015, from the Beijing Jiaotong University of the Beijing, China. She is currently pursuing the Ph.D. degree with Institute of Information Science, Beijing Jiaotong University, Beijing, China. Her research interests include semantic segmentation and object detection.

Yunchao Wei is currently a Postdoctoral Researcher in Beckman Institute at the University of Illinois at Urbana-Champaign, working with Prof. Thomas Huang. He received his Ph.D. degree from Beijing Jiaotong University in 2016, advised by Prof. Yao Zhao. He received Excellent Doctoral Dissertation Awards of Chinese Institute of Electronics (CIE) and Beijing Jiaotong University in 2016, the Winner prize of the object detection task (1a) in ILSVRC 2014, the Runner-up prizes of all the video object detection tasks in ILSVRC 2017, the Winner Prizes of all human parsing tracks in the 2nd LIP challenge. His current research interest focuses on computer vision techniques for large-scale data analysis. Specifically, he has done work in weakly- and semi-supervised object recognition, multi-label image classification, video object detection and multi-modal analysis.



Jianan Li is currently working toward the Ph.D. degree at the School of Optoelectronics, Beijing Institute of Technology, Beijing, China. His research interests mainly include computer vision and real-time image/video processing.

Shikui Wei received the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), Beijing, China, in 2010. During his Ph.D., he visited Media Lib at Nanyang Technological University in Singapore as a jointPhD student from 2008 to 2010. From 2010 to 2011, he worked as a postdoctoral researcher in School of Computer Engineering at Nanyang Technological University in Singapore. He is currently a full Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include computer vision, multimedia content analysis, machine learning. He is currently a Professor with the Institute of Information Science, Beijing Jiaotong University. His research interests include computer vision, image/video analysis and retrieval, and copy detection.