Dual Adversarial Networks for Zero-shot Cross-media Retrieval

Jingze Chi and **Yuxin Peng***
Institute of Computer Science and Technology, Peking University, Beijing, China
pengyuxin@pku.edu.cn

Abstract

Existing cross-media retrieval methods usually require that testing categories remain the same with training categories, which cannot support the retrieval of increasing new categories. Inspired by zero-shot learning, this paper proposes zero-shot cross-media retrieval for addressing the above problem, which aims to retrieve data of new categories across different media types. It is challenging that zero-shot cross-media retrieval has to handle not only the inconsistent semantics across new and known categories, but also the heterogeneous distributions across different media types. To address the above challenges, this paper proposes Dual Adversarial Networks for Zero-shot Cross-media Retrieval (DANZCR), which is the first approach to address zero-shot cross-media retrieval to the best of our knowledge. Our DANZCR approach consists of two GANs in a dual structure for common representation generation and original representation reconstruction respectively, which capture the underlying data structures as well as strengthen relations between input data and semantic space to generalize across seen and unseen categories. Our DANZCR approach exploits word embeddings to learn common representations in semantic space via an adversarial learning method, which preserves the inherent cross-media correlation and enhances the knowledge transfer to new categories. Experiments on three widely-used cross-media retrieval datasets show the effectiveness of our approach.

1 Introduction

Multimedia data has been an important part of our life including image, text, video, audio and other types, and human cognition of outside world is through the fusion of multiple sensory channels, such as vision and audition. Thus, it is quite important for artificial intelligence to effectively process and utilize multimedia data. In that case, many multimedia analysis tasks have been widely studied [Peng *et al.*, 2017; Xian *et al.*, 2017; Li *et al.*, 2018], and cross-media retrieval becomes a research hotspot, which retrieves data of different media types using a query of any media type [Peng *et al.*, 2017]. Different from single-media retrieval, the key challenge of cross-media retrieval is that the distributions and representations of different media types are inconsistent, which makes it hard to measure the similarity among cross-media data. Moreover, the collecting and labeling of cross-media data are heavily labor-consuming and time-consuming for numerous and dynamically increasing categories in the real world. Cross-media retrieval methods usually require training data of all the retrieval categories to train the retrieval models, which lack of extensibility to retrieve data of ***new categories***.

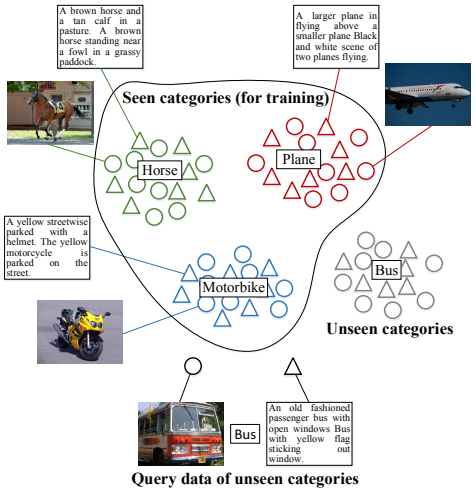

Figure 1: The brief illustration of zero-shot cross-media retrieval.

Zero-shot learning [Larochelle *et al.*, 2008] is a promising solution for recognizing new categories with limited training categories. Inspired by this, we propose ***zero-shot cross-media retrieval*** by extending zero-shot learning to the cross-media scenario, which aims to retrieve data of new categories across different media types. Different from existing zero-shot learning for single-media analysis, zero-shot cross-media retrieval intends to achieve retrieval across multiple media types in zero-shot scenario where there are no over-

*corresponding author.

laps between categories of training and testing data. A brief illustration of zero-shot cross-media retrieval is shown in Figure 1, in which categories of query data are never seen in the training. We denote known categories in the training as seen categories, and new categories not included in the training as unseen categories. It is very challenging to handle not only the inconsistent semantics across seen and unseen categories but also the heterogeneous distributions of cross-media data.

For addressing the above problems, this paper proposes a Dual Adversarial Networks for Zero-shot Cross-media Retrieval (DANZCR) approach, which exploits word embeddings as semantic space, and transforms data of different media types into semantic space via an adversarial learning framework. More specifically, the proposed approach establishes generative adversarial networks (GANs) in a dual structure, where the *forward GAN* learns from the input image and text to generate common representations in semantic space; and the *reverse GAN* uses the generated common representations to reconstruct the input image and text for preserving the original data structure. Our DANZCR approach performs zero-shot learning and correlation learning at the same time, which can generate common representations to conduct zero-shot cross-media retrieval.

The main contributions of this paper can be concluded as follows:

- ***A dual GANs architecture*** is proposed for zero-shot cross-media retrieval, which consists of forward GAN and reverse GAN. The dual GANs collaboratively promote each other, which capture the underlying data structures, as well as strengthen relations between input data and semantic space to generalize across seen and unseen categories.
- ***An adversarial training method*** is proposed for zero-shot cross-media retrieval, which learns common representations by discriminating the generated common representations from which media types and categories to preserve the inherent cross-media correlation. Word embeddings are exploited for generating common representations to model semantic information via the adversarial training process, which enhances the knowledge transfer to unseen categories.

Comprehensive experimental results show the effectiveness of our proposed approach for zero-shot cross-media retrieval on three widely-used cross-media retrieval datasets: Wikipedia, Pascal Sentence and NUS-WIDE.

2 Related Work

2.1 Cross-media Retrieval

Cross-media retrieval aims to measure similarities among data of different media types, which are of different feature spaces. Most existing methods focus on common representation learning, which can be divided into traditional methods and Deep Neural Networks (DNN) based methods. Traditional methods mainly learn linear projections for different media types. Canonical correlation analysis (CCA) [Rasiwasia *et al.*, 2010] is proposed to learn cross-media common representation by maximizing the pairwise correlation. Joint representation learning (JRL) [Zhai *et al.*, 2014] is proposed to make use of semi-supervised regularization and semantic information, which can jointly learn common representation projections for up to five media types. Kang et al. [Kang *et al.*, 2015] use a local group based priori to exploit popular block based features and jointly learn basis matrices for different media types. Recently, DNN-based cross-media retrieval has become an active research topic. Wang et al. [Wang *et al.*, 2017] propose an adversarial learning method to learn common subspace, which is implemented as an interplay between feature projector and modality classifier. Wei et al. [Wei *et al.*, 2017] propose to use CNN model pre-trained on ImageNet as the feature extractor for images, and show the effectiveness of CNN feature in cross-media retrieval. However, existing methods usually require that testing categories are the same as training categories, which cannot satisfy the requirement of extensibility in real-world applications.

2.2 Zero-shot Learning

Zero-shot learning aims to address the increasing difficulty caused by collecting data for the large number of categories. Most existing methods take advantage of external knowledge, such as encyclopedias and attributes, to improve the scalability for recognizing the categories beyond training set. Textual descriptions in Wikipedia articles are used as external knowledge in [Elhoseiny *et al.*, 2014] to learn visual classifiers for images. Wu et al. [Wu *et al.*, 2014] make use of information from search engine to address zero-shot event detection. Farhadi et al. [Farhadi *et al.*, 2009] describe objects with manually annotated attributes to classify unseen objects. In [Ba *et al.*, 2016], the attributes are extracted for unseen image classification from textual descriptions. However, these methods are mostly designed for single-media scenario, which cannot support cross-media retrieval. Our DANZCR approach exploits word embeddings from trained Natural Language Processing model as external knowledge, and performs zero-shot learning and correlation learning at the same time to conduct zero-shot cross-media retrieval. To the best of our knowledge, our proposed DANZCR is the first approach to address the problem of zero-shot cross-media retrieval.

2.3 Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) have achieved impressive results after the work by Goodfellow et al [Goodfellow *et al.*, 2014]. Most of the existing GANs-based works focus on the generative problem to generate new data. Radford et al. [Radford *et al.*, 2015] propose deep convolutional generative adversarial networks (DCGANs) to generate images from a noise input by using deconvolutions. Mehdi et al. [Mirza and Osindero, 2014] propose conditional GANs (cGANs) to add a condition on both the generator and discriminator to generate data instead of uncontrollable noise input. Ledig et al. [Ledig *et al.*, 2017] propose super-resolution generative adversarial networks (SRGANs) with a perceptual loss function, which consists of an adversarial loss and a content loss. Recent works on image translation, such as DualGAN [Yi *et al.*, 2017] and CycleGAN [Zhu *et al.*, 2017], explore GANs in a dual task to transform image style between two image domains. However, the aforementioned works

only deal with the generative problem from one media type to another through one pathway network structure, which cannot model the distribution over the cross-media input. Inspired by works on image translation, we consider zero-shot cross-media retrieval as a domain transform problem between input space and semantic space. We utilize GANs as basic modules in a dual structure for each media type, and generate common semantic representations in a multi-pathway network for cross-media data.

3 Our DANZCR Approach

3.1 Problem Definition

We take image and text media types as examples to give the formulation of zero-shot cross-media retrieval. The cross-media dataset is represented as $D = \{D_{UD}, D_{UQ}, D_{SD}, D_{SQ}\}$, where D_{UD}, D_{UQ}, D_{SD} and D_{SQ} denote unseen category data set, unseen category query set, seen category data set, and seen category query set respectively. For each subset, $D_{UD} = \{i_n, t_n, l_n\}_{n=1}^{N_{UD}}$, where i_n, t_n and l_n denote n-th instance of image, text and corresponding category label respectively, and N_{UD} is the total number of the instances in D_{UD}. Similarly, $D_{UQ} = \{i_n, t_n, l_n\}_{n=1}^{N_{UQ}}$, $D_{SD} = \{i_n, t_n, l_n\}_{n=1}^{N_{SD}}$, and $D_{SQ} = \{i_n, t_n, l_n\}_{n=1}^{N_{SQ}}$. Notably, there is no overlap categories between seen category set $\{D_{SD}, D_{SQ}\}$ and unseen category set $\{D_{UD}, D_{UQ}\}$. Our goal is to retrieve relevant instances of different media types. Specifically, during the training stage, only the seen category data set D_{SD} can be used to train the method. In the testing stage, to distinguish between zero-shot scenario and conventional scenario, two retrieval tasks are introduced: zero-shot cross-media retrieval takes queries from D_{UQ} to retrieve relevant cross-media data in D_{UD}, and conventional cross-media retrieval use queries from D_{SQ} and search their relevant instances in D_{SD}.

3.2 Architecture of DANZCR

The overview illustration of our DANZCR approach is shown as Figure 2. A pair of GANs: forward and reverse GANs in a dual structure are adopted to form two parallels for both image and text. For each media type, the data original representations are first extracted and the forward GAN generates common representations from the original representations, and then the reverse GAN transforms the generated representations to data original representations as a dual process. Each GAN consists of a generative model and a discriminative model. We introduce the detailed network architectures of the forward and reverse GAN respectively as follows:

1) Forward GAN: The forward GAN is designed to generate semantic common representations with category word embeddings as supervised information, in order to model correlation across different medias and categories. G_{FI} and D_{FI} denote forward generative model and discriminative model of forward GAN for image, and G_{FT} and D_{FT} are for text.

The forward generative models generate common representations from image and text original representations, which are built with several fully-connected layers. For image instance i_n, the image original representation is defined as f_n^i and common representation is s_n^i, f_n^t and s_n^t are for text instance t_n. So there are equations $s_n^i = G_{FI}(f_n^i)$ and $s_n^t = G_{FT}(f_n^t)$.

The forward discriminative models are also made of several fully-connected layers for image and text. Both of them attempt to discriminate which media types and categories the generated common representations belong to. The forward discriminative models take concatenation of generated common representation and data original representation as input. The output is a single value to predict the generated common representation is real or not, as well as discriminate the semantic correlation between common representation and data representation. D_{FI} tries to determine the word embedding for corresponding category s_n^l as the real data, while generated common representation s_n^i, generated common representation of corresponding text s_n^t and word embedding for irrelevant category $s_n^{l'}$ as the fake data. D_{FT} is similar to D_{FI}.

The objective function of forward GAN can be defined as:

$$\begin{aligned} L_{GF} = E_{i,t,l\sim P_{i,t,l}}[&log(D_{FI}(s_n^l, f_n^i)) + log(D_{FT}(s_n^l, f_n^t)) \\ &+ log(1 - D_{FI}(s_n^i, f_n^i)) + log(1 - D_{FT}(s_n^t, f_n^t)) \\ &+ log(1 - D_{FI}(s_n^t, f_n^i)) + log(1 - D_{FT}(s_n^i, f_n^t)) \\ &+ log(1 - D_{FI}(s_n^{l'}, f_n^i)) + log(1 - D_{FT}(s_n^{l'}, f_n^t))] \end{aligned} \tag{1}$$

2) Reverse GAN: The reverse GAN is designed to reconstruct the original representations of input data, which makes common representations to preserve the data original structure, as well as strengthens relations between input data and semantic space. G_{RI} and D_{RI} denote reverse generative model and discriminative model of reverse GAN for image, and G_{RT} and D_{RT} for text.

The reverse generative models are built with several fully-connected layers , which learn from common representations of the forward GAN to generate the reconstruction representations of input data original representations. The reconstruction representations are defined as r_n^i and r_n^t for image and text respectively. So there are equations $r_n^i = G_{RI}(G_{FI}(f_n^i)) = G_{RI}(s_n^i)$ and $r_n^t = G_{RT}(G_{FT}(f_n^t)) = G_{RT}(s_n^t)$.

The reverse discriminate models of image and text are also constructed by several fully-connected layers. They take the data representation as input and output a single value to predict the data representation is real or not. D_{RI} attempts to distinguish image representation f_n^i as real with the generated reconstruction representation r_n^i as fake. D_{RT} is similar for text.

The objective function of reverse GAN is:

$$\begin{aligned} L_{GR} = &E_{i\sim P_i}[log(D_{RI}(f_n^i)) + log(1 - D_{RI}(r_n^i))] \\ &+ E_{t\sim P_t}[log(D_{RT}(f_n^t)) + log(1 - D_{RT}(r_n^t))] \end{aligned} \tag{2}$$

However, for zero-shot cross-media retrieval, there are usually a few data of limited known categories for training. If there are large amounts of data, the model has the capacity to learn transform mode from input data to common representation in semantic space, which is unrealistic and meaningless.

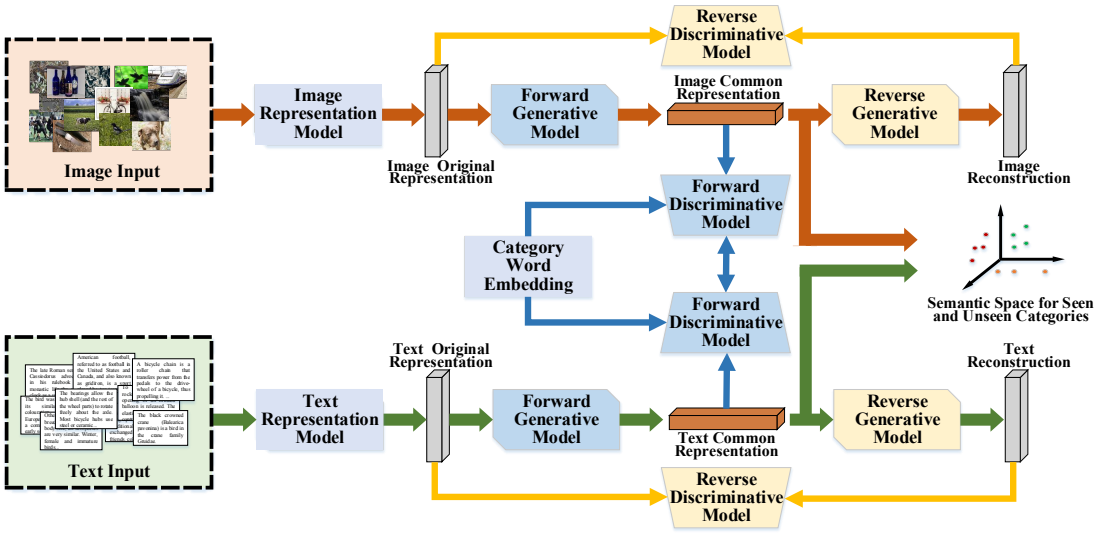

Figure 2: The overview illustration of our DANZCR approach.

So to further enforce the correlation learning, a recovery objective function via the l_2 norm is proposed as follows:

$$L_{RE} = E_{i,l\sim P_{i,l}}[||s_n^i - s_n^l||_2] + E_{t,l\sim P_{t,l}}[||s_n^t - s_n^l||_2] + E_{i\sim P_i}[||r_n^i - f_n^i||_2] + E_{t\sim P_t}[||r_n^t - f_n^t||_2] \tag{3}$$

According to the above definitions, our DANZCR can be trained by jointly learning of two dual GANs for both image and text. The full objective function of DANZCR approach is:

$$L = L_{GF} + L_{GR} + L_{RE} \tag{4}$$

3.3 Training Procedure

With the defined objective functions, the generative model and discriminative model can play a minimax game with each other, which are trained iteratively in an adversarial process. The parameters of generative model are fixed during the discriminative model training stage and vice versa. The optimization procedure of our approach are presented as follows:

1) Common representation generation: The image and text original representations are extracted and then the common representations are generated from the forward generative models.

2) Optimizing forward discriminative models: The forward discriminative models are trained by ascending stochastic gradient for image and text respectively as follows:

$$\nabla\theta_{D_{FI}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{FI}(s_n^l, f_n^i)) + log(1 - D_{FI}(s_n^i, f_n^i)) + log(1 - D_{FI}(s_n^t, f_n^i)) + log(1 - D_{FI}(s_n^{l'}, f_n^i))] \tag{5}$$

$$\nabla\theta_{D_{FT}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{FT}(s_n^l, f_n^t)) + log(1 - D_{FT}(s_n^t, f_n^t)) + log(1 - D_{FT}(s_n^i, f_n^t)) + log(1 - D_{FT}(s_n^{l'}, f_n^t))] \tag{6}$$

where N is the number of an instance batch.

3) Representation reconstruction: The reconstruction representations are generated by the reverse generative models from generated common representations.

4) Optimizing reverse discriminative models: The stochastic gradient equations can be defined to optimize the reverse discriminative models for image and text respectively as:

$$\nabla\theta_{D_{RI}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{RI}(f_n^i) + log(1 - D_{RI}(r_n^i)] \tag{7}$$

$$\nabla\theta_{D_{RT}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{RT}(f_n^t) + log(1 - D_{RT}(r_n^t)] \tag{8}$$

5) Optimizing forward generative models: The equations to optimize the forward generative models for image and text are as follows:

$$\nabla\theta_{G_{FI}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{FI}(s_n^i, f_n^i)) + log(D_{FI}(s_n^t, f_n^i)) + \lambda_F||s_n^i - s_n^l||_2] \tag{9}$$

$$\nabla\theta_{G_{FT}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{FT}(s_n^t, f_n^t)) + log(D_{FT}(s_n^i, f_n^t)) + \lambda_F||s_n^t - s_n^l||_2] \tag{10}$$

5) Optimizing forward and reverse generative models: The forward and reverse generative models are updated by descending the stochastic gradient as follows:

$$\nabla\theta_{G_{RI},G_{FI}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{RI}(r_n^i)) + \lambda_R||r_n^i - f_n^i||_2] \tag{11}$$

$$\nabla\theta_{G_{RT},G_{FT}} = \frac{1}{N}\sum_{n=1}^{N}[log(D_{RT}(r_n^t)) + \lambda_R||r_n^t - f_n^t||_2] \tag{12}$$

where the parameter λ_F and λ_R control the relative contribution of the recovery objective terms.

3.4 Implementation

We adopt TensorFlow [1] to implement our model with a base learning rate 1^{-4} and dropout probability 0.9. The parameters λ_F and λ_R are set to 1^{-2}. The word embeddings for categories are the 300-dimensional features extracted by Word2Vec model [Mikolov *et al.*, 2013], which are pre-trained on Google News. The image representation model has the same configuration with 19-layer VGGNet [Simonyan and Zisserman, 2014] and 4,096 dimensional feature vector from fc7 layer is extracted as the image original representation. The 300-dimensional text original representation is extracted by Doc2Vec model [Le and Mikolov, 2014],which is also pre-trained on Google News. Then the forward generative models with three fully-connected layers are adopted for both image and text to generate common representations with each layer following a ReLU layer and a dropout layer except the last. The number of hidden units are 4,096, 4,096 and 300. The reverse generative models of both image and text are composed of three fully-connected layers to reconstruct image and text representations, with the 4,096 hidden units for the first two layers. The forward and reverse discriminative model have similar structure of three fully-connected layers with 4,096, 2,048 and 1 hidden units, which project the input data into a single value as a predict score to distinguish real or fake.

4 Experiments

4.1 Dataset

Wikipedia dataset [Rasiwasia *et al.*, 2010] is widely used for cross-media retrieval evaluation. It is based on "featured articles" in Wikipedia, which contains 2,866 image/text pairs with 10 high-level semantic categories. The dataset is randomly split into two parts: 2,173 pairs are selected as training set and 693 pairs are selected as testing set.

Pascal Sentence dataset [Farhadi *et al.*, 2010] is selected from 2008 PASCAL development kit, which contains 1,000 image/text pairs organized into 20 categories. Each image instance has 5 sentences as description. There are 800 image/text pairs selected as training set and 200 image/text pairs for testing set.

NUS-WIDE dataset [Chua *et al.*, 2009] consists of about 270,000 images with their tags categorized into 81 categories. Only the images exclusively belonging to one of the 10 largest categories in NUS-WIDE dataset are selected for experiments following [Zhuang *et al.*, 2013], and each image along with its corresponding tags is viewed together as an image/text pair with unique category label. Finally, there are about 70,000 image/text pairs, where training set consists of 42,941 pairs and testing set consists of 28,661 pairs.

For all three datasets, we conduct two cross-media retrieval tasks to evaluate our approach comprehensively: zero-shot retrieval and conventional retrieval according to the definition in Section 3.1. Inspired by [Xian *et al.*, 2017; Liu *et al.*, 2017], we employ a new dataset split setting based on the original split settings, which is shown in Figure 3. The training set and testing set in the original split setting are further divided into two sets: seen category set and unseen category set respectively, and each set includes 50% categories.

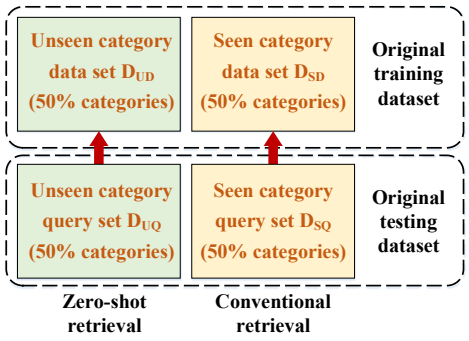

Figure 3: The dataset splitting of two retrieval tasks.

It is noted that input features and evaluation protocol in this paper are different with the compared methods. For the original results of the compared methods, the features are mostly obtained with fine-tuning models, and evaluation protocol is designed to test only on seen categories. While in this paper, we use input features without fine-tuning and the evaluation protocol is designed to test on both seen and unseen categories. For fair and objectively comparison, we directly re-run the source codes of all compared methods provided by their authors.

4.2 Evaluation Metrics

Taking Image→Text as an example, we take each image as a query, and measure the cosine distance between the common representations of the query image and all texts. Finally, we get a ranking list according to the distances and then compute the mean average precision (MAP) to evaluate the retrieval results. The MAP scores are computed as all queries' mean of average precision (AP), and AP is computed as:

$$AP = \frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k} \times rel_k \tag{13}$$

where R denotes relevant item number in testing set (according to the category label in our experiments), R_k denotes the relevant item number in top k results, n denotes the testing set size, and $rel_k = 1$ means the k-th result is relevant, and 0 otherwise.

4.3 Experimental Analysis

Among the compared methods, LGCFL, JRL and CCA are traditional methods, while Deep-SM and ACMR are DNN-based methods. The MAP scores for zero-shot cross-media retrieval are shown in Table 1, in which our DANZCR approach achieves the best results on the three datasets. We can observe that for the compared methods there is no obvious gap between results of DNN-based methods and traditional methods, and some traditional methods are even higher than DNN-based methods. This is because testing categories of

[1] www.tensorflow.org

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	**our DANZCR**	**0.297**	**0.287**	**0.292**
	deep-SM	0.265	0.258	0.262
	ACMR	0.276	0.262	0.269
	LGCFL	0.261	0.258	0.260
	JRL	0.264	0.266	0.265
	CCA	0.238	0.236	0.237
Pascal Sentences dataset	**our DANZCR**	**0.334**	**0.338**	**0.336**
	deep-SM	0.276	0.251	0.264
	ACMR	0.306	0.291	0.299
	LGCFL	0.273	0.258	0.266
	JRL	0.298	0.283	0.291
	CCA	0.207	0.183	0.195
NUS-WIDE dataset	**our DANZCR**	**0.416**	**0.469**	**0.443**
	deep-SM	0.401	0.414	0.408
	ACMR	0.407	0.425	0.416
	LGCFL	0.396	0.422	0.409
	JRL	0.401	0.449	0.425
	CCA	0.400	0.397	0.399

Table 1: MAP scores of our approach and compared methods for zero-shot cross-media retrieval.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	**our DANZCR**	**0.672**	**0.887**	**0.778**
	deep-SM	0.674	0.872	0.773
	ACMR	0.672	0.865	0.769
	LGCFL	0.510	0.586	0.548
	JRL	0.522	0.604	0.563
	CCA	0.261	0.267	0.264
Pascal Sentences dataset	**our DANZCR**	**0.737**	**0.868**	**0.803**
	deep-SM	0.728	0.841	0.785
	ACMR	0.726	0.756	0.741
	LGCFL	0.592	0.638	0.615
	JRL	0.636	0.677	0.657
	CCA	0.214	0.183	0.199
NUS-WIDE dataset	**our DANZCR**	**0.727**	**0.709**	**0.718**
	deep-SM	0.680	0.667	0.674
	ACMR	0.604	0.702	0.653
	LGCFL	0.459	0.529	0.494
	JRL	0.480	0.616	0.548
	CCA	0.432	0.438	0.435

Table 2: MAP scores of our approach and compared methods for conventional cross-media retrieval.

Method	Task		
	Image→Text	Text→Image	Average
our DANZCR	**0.334**	**0.338**	**0.336**
DANZCR-Single	0.312	0.307	0.310
DANZCR-NoDis	0.314	0.320	0.317

Table 3: Baseline experiments on Pascal Sentences dataset for zero-shot cross-media retrieval.

Method	Task		
	Image→Text	Text→Image	Average
our DANZCR	**0.737**	**0.868**	**0.803**
DANZCR-Single	0.728	0.850	0.789
DANZCR-NoDis	0.711	0.843	0.777

Table 4: Baseline experiments on Pascal Sentences dataset for conventional cross-media retrieval.

zero-shot retrieval are not included in the training, which may limit the learning ability of DNN-based methods. The MAP scores for conventional cross-media retrieval are shown in Table 2, and our proposed DANZCR approach also achieves the best retrieval accuracy among all the compared methods. For conventional cross-media retrieval, the DNN-based methods are all higher than traditional methods.

Results of conventional cross-media retrieval are obviously higher than zero-shot cross-media retrieval for all methods. Because zero-shot cross-media retrieval is more challenging in which categories of testing data are never seen during the training. Our DANZCR achieves the best accuracy compared with other methods for both zero-shot and conventional cross-media retrieval, for exploring the external knowledge from the word embeddings of different categories via a dual GANs structure, which can learn more discriminative common representations and ensure the generalization ability across seen and unseen categories to support zero-shot cross-media retrieval.

To further present the effectiveness for each component in our proposed approach, three kinds of baseline experiments are conducted on Pascal Sentences dataset, which are shown in Table 3 and 4. "DANZCR-Single" means only the forward GAN is used without the reverse GAN. "DANZCR-NoDis" means discriminating terms are not used in the adversarial training method, which refers to removing the items $log(1 - D_{FI}(s_n^t, f_n^i))$, $log(1 - D_{FT}(s_n^i, f_n^t))$, $log(1 - D_{FI}(s_n^{l'}, f_n^i))$ and $log(1 - D_{FT}(s_n^{l'}, f_n^t))$ in Equation 1. We can observe that our DANZCR approach obtains higher accuracy than "DANZCR-Single", which demonstrates that reverse GAN model can effectively preserve semantic consistency for each media and collaboratively promote forward GAN to generalize across seen and unseen categories. Results of "DANZCR-NoDis" decline compared with DANZCR, which indicates that discrimination information across different media types and categories are both beneficial to learn more discriminative semantic common representations.

5 Conclusion

This paper has proposed a Dual Adversarial Networks for Zero-shot Cross-media Retrieval (DANZCR) approach, which performs zero-shot learning and correlation learning at the same time to generate common representations for zero-shot cross-media retrieval. The dual GANs architecture consists of forward GAN and reverse GAN to collaboratively promote each other, which captures the underlying data structures, as well as strengthens relations between input data and semantic space to generalize across seen and unseen categories. Besides, the adversarial training method discriminates the generated common representations from which media types and categories to preserve the inherent cross-media correlation, which makes the generated common representations model semantic information and enhances the knowledge transfer to unseen categories.

The future work lies in the following aspects: Firstly, external knowledge will be effectively utilized to build connections between seen and unseen categories. Secondly, a large number of unlabeled data can be exploited in the future to form an unsupervised model.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, pages 4247–4255, 2016.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, pages 1–9, 2009.

[Elhoseiny *et al.*, 2014] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, pages 2584–2591, 2014.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[Farhadi *et al.*, 2010] Ali Farhadi, Mohsen Hejrati, Mohammad Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Kang *et al.*, 2015] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia (TMM)*, 17(3):370–381, 2015.

[Larochelle *et al.*, 2008] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, pages 646–651, 2008.

[Le and Mikolov, 2014] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 105–114, 2017.

[Li *et al.*, 2018] Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. Gla: Global–local attention for image description. *IEEE Transactions on Multimedia*, 20(3):726–737, 2018.

[Liu *et al.*, 2017] Ruoyu Liu, Yao Zhao, Liang Zheng, Shikui Wei, and Yi Yang. A new evaluation protocol and benchmarking results for extendable cross-media retrieval. *arXiv preprint arXiv:1703.03567*, 2017.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Peng *et al.*, 2017] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2017.

[Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260, 2010.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Wang *et al.*, 2017] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MM*, pages 154–162, 2017.

[Wei *et al.*, 2017] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Transactions on Cybernetics (TCYB)*, 47(2):449–460, 2017.

[Wu *et al.*, 2014] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, pages 2665–2672, 2014.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 3077–3086, 2017.

[Yi *et al.*, 2017] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017.

[Zhai *et al.*, 2014] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 24(6):965–978, 2014.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017.

[Zhuang *et al.*, 2013] Yueting Zhuang, Yanfei Wang, Fei Wu, Yin Zhang, and Weiming Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.