Applied Intelligence (2023) 53:4257–4267
https://doi.org/10.1007/s10489-022-03653-7

Dual discriminant adversarial cross-modal retrieval

Pei He[1] · Meng Wang[2] · Ding Tu[2] · Zhuo Wang[1]

Accepted: 19 April 2022 / Published online: 6 June 2022

Abstract

In order to improve the accuracy of cross-modal retrieval tasks and achieve flexible retrieval between different modalities, we propose a Dual Discriminant Adversarial cross-modal Retrieval (DDAC) method in this paper. First, DDAC integrates adversarial learning and minimization of feature projection distances and introduces label information in it. It can eliminate the same semantic heterogeneity between modalities while maintaining the distinguishability of different semantic features between modalities. Then, cosine distance is used to minimize and maximize the inter-modal distance of features with the same and different labels respectively to solve the inter-modal discrimination problem. Different from the general method, DDAC performs dual discrimination in the label space and solves the intra-modal discrimination problem from two perspectives of probability distribution and distance. Extensive experiments carried out on three public datasets validate that the proposed DDAC outperforms the state-of-the-art methods.

Keywords Dual discriminant · Inter-modal consistency · Cross-modal retrieval · Adversarial learning

1 Introduction

In recent decades, with the development of advanced technology, multimedia data on various search engines and social media shows the trend of explosive growth. Among them, the growth of different types of media data, such as text, image, audio, and video, makes cross-modal retrieval more and more important in practical applications [1]. Different from single-modal retrieval [6, 12], the purpose of cross-modal retrieval is to search for related samples in another modality based on the query samples in one modality. For example, given an image, to retrieve a text description containing the same object or topic. However, because each modality has different data representation forms, the similarity comparison between different types of data cannot be conducted directly.

The core of cross-modal retrieval is to find a subspace where the similarity of different modal data can be directly compared. For example, Canonical Correlation Analysis(CCA) [2] is one of the representative subspace learning methods, which reflects the linear correlation between two sets of heterogeneous variables. However, in general, shallow methods tend to ignore the underlying features, and due to the complexity of data, the retrieval effect of linear projection modeling is not ideal. The development of deep learning has also had a great impact on cross-modal retrieval, its powerful automatic feature extraction capability has been widely used in cross-modal retrieval tasks [7, 9, 10]. Andrew et al. introduced deep learning into CCA and proposed Deep Canonical Correlation Analysis (DCCA) [7], which builds two multi-layer Deep networks to learn complex nonlinear projections and maximize the Correlation of common representations after projections. Wei et al. fine-tuned the pre-trained CNN model and proposed deep Semantic Matching [34]. Deep-SM uses different loss functions for different target datasets, and uses fine-tuned CNN and trained fully connected neural

✉ Meng Wang
mwang007@163.com

Pei He
hepei139@foxmail.com

Ding Tu
tuding_NEAU@163.com

Zhuo Wang
wz1181620976@163.com

[1] College of Science, Guangxi University of Science and Technology, Liuzhou, 545000, China

[2] Tus College of Digit, Guangxi University of Science and Technology, Liuzhou, 545000, China

networks to project images and texts into high-dimensional homogeneous semantic Spaces.

Motivation. In the use of label information, ACMR [13] uses a label classifier to minimize the probability distribution of the real label and the predicted label to ensure the discriminability between modalities. DSCMR [14] uses the F norm to minimize the distance between the real label and the predicted label. We believe that combining these two methods in an appropriate way can better ensure the discriminativeness between modalities. In addition, although adversarial learning is widely used, it usually confuses the discriminant by the projection of unknown features to eliminate the heterogeneity of all features between the modalities. In fact, because the ultimate purpose of cross-modal retrieval is to make inter-modal feature projection under the same semantics closest, we believe that the difference of inter-modal feature projection with different semantics can be retained, and only the heterogeneity of feature projection under the same semantics can be eliminated.

Our Method. In this paper, we proposes a Dual Discriminant Adversarial cross-modal Retrieval (DDAC) method for cross-modal Retrieval. DDAC combines adversarial learning with the minimization of feature projection distance and can eliminate the heterogeneity between modalities. In the adversarial learning part, we add a rejector R to the discriminator D. Its function is to restrain the projection of unknown features to confuse the discriminator D, so as to achieve the purpose of maintaining the distinguishability of the same semantic features among the modalities. At the same time, in order to maintain consistency with adversarial learning, we only minimize the feature projection distances that have the same semantics between modalities. Then, DDAC uses cosine distance to expand and shrink the feature representation distance between modalities with the same semantics and different semantics, to ensure the minimum discriminant loss between modalities. Finally, a label classifier with branching is used to dual predict the semantic tags of the projected items from the perspective of probability distribution and distance, to ensure the intra-modal invariance of the classification information. The main work of this paper is as follows:

- We propose a cross-modal retrieval approach to eliminate the inter-modal heterogeneity and ensure semantic consistency by making full use of label information.
- The improved adversarial learning was applied to DDAC, and the label information was introduced into the discriminant, which could eliminate the same semantic feature heterogeneity between modalities while preserving the feature distinguishability between different semantics. At the same time, the inter-modal projection loss is minimized to further break the modal gap.
- DDAC uses a linear classifier with branches to classify samples in two different forms in the label space. In this way, DDAC makes the learned common representation more distinguishable.
- Extensive experiments on three public datasets show that DDAC is superior to current state-of-the-art methods

2 Related work

2.1 Supervised learning

In general, the supervised methods [5, 9, 20] use label information to distinguish different categories of representations in the common space, or use semantic correlation of tags to mine associations between multi-modal data. Li et al. [16] uses label information to train label semantic network as the supervision network of other sub-networks. He et al. [8] proposed a fine-grained cross-media learning method. In order to obtain better representation, it ensures the discriminance of subcategories features, compactness of the same subcategory features and sparsity of different subcategories through three constraints. In [14], Zhen et al. proposed Deep Supervised Cross-modal Retrieval (DSCMR), which minimizes the discrimination loss of samples in label space and public representation space and maintains the differences between samples of different semantic categories. At the same time, the weight-sharing strategy is used to deepen the correlation between the inter-modal features. However, the general supervised methods [26–28] including the above methods only use label information from one perspective, which may cause omissions. Therefore, we consider comprehensively utilizing the label information from different perspectives to improve the intra-modal differentiability of features.

2.2 Adversarial learning

In addition, how to bridge the semantic gap between modalities has been a problem to be solved in recent years, and adversarial learning is one of the more effective methods [13, 15]. In [13], Wang et al. proposed the Adversarial Cross-Modal Retrieval (ACMR) for the first time based on the idea of Adversarial learning. The modality classifier is constructed by ACMR to distinguish different modalities, and then the modality classifier is confused by the feature projection. The modal consistency problem is tried to be solved by the process of adversarial learning. Peng et al. constructed a cross-modal GAN system called cross-modal Generative Adversarial Networks (CM-GANS) [15]. The process of adversarial learning is realized by constructing the cross-mode convolution autoencoder as the generator, and then using the two discriminant models to discriminate between the modes within the modes simultaneously. Chen et al. also solved the semantic gap between modalities by introducing adversarial learning and combining it with Shannon Information Theory [4]. When modality classification is carried out, the information entropy is maximized. The SCH-GAN model proposed by Zhang et al. [32] is a semi-supervised cross-modal hashing learning method based on generative adversarial network and uses a reinforcement learning-based algorithm to drive the training of model. Adversarial learning has been widely used in current approaches [16, 17, 31, 33] because of its effectiveness.

DDAC double discriminates labels in the label space from the perspectives of probability distribution and distance, which can make full use of the label information to guide the model to effectively distinguish different categories of features. In addition, we improved the discriminator in adversarial learning by adding a rejector. While using adversarial learning to eliminate the heterogeneity gap, we could further solve the problem of inter-modal discriminant, thus improving the accuracy and stability of cross-modal retrieval.

3 The proposed DDAC

3.1 Problem formulation

In this section, we define a number of symbols. Without losing generality, We assume that there are n image-text pairs. The input of image modality is denoted as $X = \{x_i\}_{i=1}^n$, Where, x_i is the feature vector of the ith image. Again, the input of image modality is denoted as $Y = \{y_i\}_{i=1}^n$, Where, y_i is the feature vector of the text description of the ith image. In addition, the lables of these n image and text pairs are denoted as $L = \{l_i\}_{i=1}^n$, where , $l_i = \{l_{i1}, l_{i2}, \ldots, l_{ic}\} \in R^c$, c is the number of categories.

Since image feature X and text feature Y are different types of statistics and follow unknown distributions, they cannot be directly compared in cross-modal retrieval. Therefore, we need to find a common space S, so that the features of images and texts can be directly compared. The feature projection of the image is $f^X = f_X(X; \theta_X) \in R^d$,and the feature projection of the text is $f^Y = f_Y(Y; \theta_Y) \in R^d$, where d is the dimension of the common space S. θ_X and θ_Y are the parameters of the two functions.

3.2 Framework of DDAC

Figure 1 shows the total framework of the DDAC method. The first step is feature extraction. Image modality features is extracted with VGG-19 [18] which is pre-trained on ImageNet. We obtain the 4096-dimensional original high-dimensional semantic feature representation of the

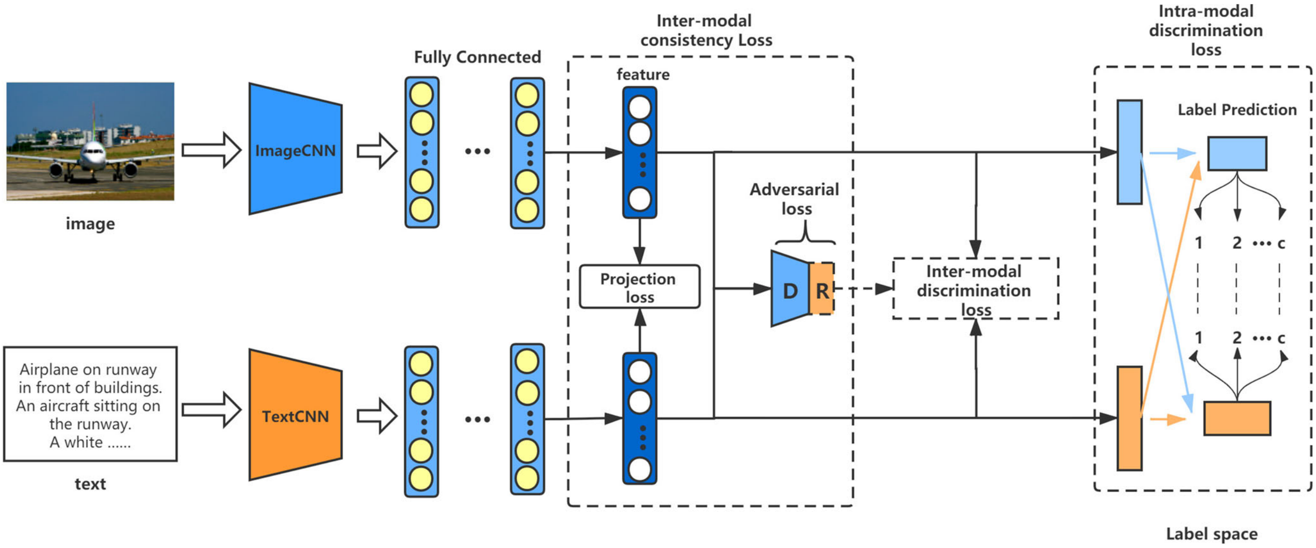


image
ImageCNN
Fully Connected
Airplane on runway in front of buildings. An aircraft sitting on the runway. A white ......
text
TextCNN
Inter-modal consistency Loss
feature
Projection loss
Adversarial loss
D
R
Inter-modal discrimination loss
Intra-modal discrimination loss
Label Prediction
1
2
c
Label space

Fig. 1 The total framework of the DDAC method

image modality from the fc-7 layer; At the same time, Word2Vec [19] model pre-trained on Google News was used to extract each text into a matrix composed of k-dimensional feature vectors, and then the original high-level semantic representation was extracted from the text feature matrix through Sentence CNN; Finally, through a set of fully-connected layer, the final image and text common representations are generated, denoted as f^X and f^Y respectively.

Next, DDAC implements cross-modal retrieval through three parts: The first part is to reduce the inter-modal heterogeneity, that is, to ensure inter-modal consistency. While minimizing the feature projection loss L_d with the same semantics, a discriminator D composed of a three-layer feedforward neural network distinguishes the unknown projection from being an image modality or a text modality. For the unknown projection, try to confuse the discriminator D. Among them, the role of the rejector R is to enable the discriminator D to better distinguish different modalities and have different semantic features. In general, in this part, we can achieve the goal of only eliminating the heterogeneity between modalities that have the same semantic features. At the same time, improve the differentiation of features by preserving the heterogeneity of different semantic features.

The second part is to use the cosine distance to scale the distance of feature projections with the same semantics and different semantics. The purpose is to minimize inter-modal discriminant loss.

The third part is to minimize the intra-modal discrimination loss in the label space. The linear classifier with branching used by DDAC can measure the similarity between the predicted label and the real label of each sample from the perspectives of probability distribution and distance, and supervise the mapping function of the sub-network so that the label information in the intra-modal data after the feature projection can be better retained. (Specific objective function construction in the three parts will be shown in detail in Section 3.3)

3.3 Objective function

3.3.1 Inter-modal consistency loss

We plan to eliminate the cross-modal gap in two ways to ensure inter-modal consistency. For samples from different modalities, we only minimize the distance between images and text features with the same label. Therefore, the following loss function is constructed as the projected loss:

$$L_d = \frac{1}{n} \|X_l - Y_l\|_F \tag{1}$$

X_l and Y_l represent image and text modality with the same label, respectively.

Next, we chose to introduce adversarial learning to further eliminate the inter-modal heterogeneous differences. DDAC constructs a discriminator D, which is composed of three layers of the feed-forward neural network and adds a rejector R. Its function is merely to eliminate the heterogeneity of feature projection with the same semantic between modalities.

In this process, unknown feature projections from different modalities try their best to confuse the discriminator so that it cannot distinguish whether the input feature is a text or an image, and the rejector inhibits the feature projection between different semantics to confuse the discriminator, to preserve the heterogeneity of inter-modal different semantic features. We'll define the adversarial loss as L_{adv}:

$$\begin{aligned} L_{adv} = & -\frac{1}{n}\sum_{i=1}^{n}\left(log\left(D\left(\sigma f_i^X;\theta_D\right)\right)\right. \\ & \left.+log\left(1-D\left(\sigma f_i^Y;\theta_D\right)\right)\right) \end{aligned} \tag{2}$$

Where, L_{adv} can be regarded as the cross entropy loss of all sample modal discrimination, $D\left(f_i^X\right)$ and $D\left(f_i^Y\right)$ represent the discriminant scores of the input image and text features, θ_D is the discriminator parameters, the rejector σ is expressed as:

$$\sigma = \begin{cases} 1, l_X = l_Y. \\ n, otherwise. \end{cases} \tag{3}$$

Where l_X and l_Y are labels of different modalities, and n is a sufficiently large and appropriate real number. At the end of the discriminator is a sigmoid function, which does not change when the labels are the same. when the labels are different, $\phi\left(f^*\right) = \frac{1}{1+e^{-nf_i^*}}$.

An appropriate n can make it easier for discriminator D to distinguish unknown feature projections with different labels. To ensure the distinguishability of features with different labels between modalities. Under the same label, the higher the discriminant score is, the more likely the input feature is to come from the image modality; the smaller the discriminant score is, the more likely it is to come from the text modality.

3.3.2 Inter-modal discrimination loss

When feature projections of different modalities belong to the same semantic category, the distance between them should be as close as possible. For feature projections of different categories among modes, the distance should be as far as possible. The above purpose is achieved by constructing the following loss function:

$$L_{imd}=\sum_{i,j=1}^{n}\Theta_{ij}cos\left(f_i^X,f_i^Y\right) \tag{4}$$

Where, Θ is the signal function. If f_i^X and f_i^Y belong to the same category , $\Theta_{ij}=1$,otherwise,$\Theta_{ij}=0$. $cos\left(\cdot\right)$ is the cosine distance of the two feature projections.

3.3.3 Intra-modal discrimination loss

To distinguish the intra-modal categories, we connect a linear classifier at the top of the subnetworks of image modality and text modality, which generates a c dimensional predictive label vector for each sample in the label space.

The first part of the label loss is as follows:

$$L_{l1}\left(P\right)=\frac{1}{n}\left(\left\|P^TX-L\right\|_F+\left\|P^TY-L\right\|_F\right) \tag{5}$$

Where, $\|\cdot\|_F$ is the Frobenius norm and P is the parameter of the linear classifier. This part of the loss function calculates the distance between the predicted label and the true label.

DDAC attempts to enhance the use of label information by adding a Softmax activation function at the end of the linear classifier described above to output the probability distribution of each semantic category. We use the probability distribution $\hat{p}$ to represent the second part of the Intra-modal discrimination loss:

$$L_{l2}\left(P\right)=-\frac{1}{n}\left(l_i\cdot\left(log\hat{p}_i\left(x_i\right)+log\hat{p}_i\left(y_i\right)\right)\right) \tag{6}$$

Where, L_{l2} represents the cross-entropy loss of semantic classification of all instances, P is the parameter of the linear classifier, l_i represents the true label of each sample, and $\hat{p}_i$ represents the probability distribution of each sample category.

3.4 Optimization

The total loss function is:

$$L_{lT}=\lambda L_{l1}+\eta L_{l2} \tag{7}$$

$$L_{imc}=\beta L_d+\gamma L_{adv} \tag{8}$$

$$L_{Total}=L_{lT}+L_{imc}+\alpha L_{imd} \tag{9}$$

Where, L_{lT} is the total intra-modal discriminant loss; L_{imc} is the total loss of inter-modal consistency,λ and η are hyper-parameters, which control the weight of different label discrimination angles; β and γ are the hyper-parameters that can control the weight of the projection loss and adversarial loss; is the hyper-parameter controlling the weight of inter-modal discrimination loss.

In this paper, the stochastic gradient descent method [21] was adopted to optimize the overall objective function, and the details of the optimization process were summarized in Algorithm 1.

Algorithm 1 Optimization process of DDAC method.

```
Input:
  The learning rate $\tau$ ,Training set $\Psi=\{x_i,y_i\}_{i=1}^{n}$,the label matrix L,batch size m,the maximal number of epochs E,hyper-parameters $\lambda,\eta,\gamma,\beta,\alpha$
Output:
  The sub-networks parameters $\theta_X$ and $\theta_Y$,Linear classifier parameters P and discriminantor $\theta_D$;
Randomly initialise the parameters;
for a =1,2,...,E do
    for b = 1,2,...,B do
        Randomly sample a batch of m image-text pairs from $\Psi$;
        Computes feature representations of images $f_i^X$ and text $f_i^Y$ ;
        Calculate the objective function equation (9)
        The stochastic gradient descent algorithm is used to calculate the gradient of the total loss function $L_{Total}$.
        Update parameters P and $\theta_D$ by backward propagation.
        Update sub-networks parameters $\theta_X$ and $\theta_Y$ by backward propagation.
    end for
end for
```

4 Experiments

4.1 Datasets and evaluation metric

Datasets: In this section, we will experiment on several datasets that are widely used for cross-modal retrieval. For a fair comparison, we exactly follow the dataset partition and feature extraction strategies from [14].

The Pascal Sentence datasets [22] includes 1000 images of 20 categories. Each image is accompanied by a document, which contains 5 sentences describing the picture. We took 800 image-text pairs from the data set as the training set, and 100 image-text pairs as the test set.

The Wikipedia datasets [29] consist of 2866 image-text pairs containing 10 semantic classes in total. We took out 2176 image-text pairs from the dataset as the training set, and 462 image-text pairs as the test set.

NUS-WIDE-TC21 datasets [3] contains 195,834 image and text pairs in 21 categories. We selected the 10 categories with the most samples, and randomly selected 8481 image-text pairs as training set and 2709 image-text pairs as test set.

Among them, the text is represented by a 1000-dimensional BOW vector.

Evaluation Metric: In this paper, we use the commonly used evaluation metrics mAP, pr-curve, and Recall@K (K=1, 3, 5) in cross-modal retrieval for evaluation. In our experiments, we compute three metrics on two different tasks: retrieving text with query image(Image2Text)and retrieving image with query text(Text2Image).

4.2 Comparison with State-of-the-art method

To prove the superiority of DDAC method, we compared the method DDAC in this paper with the 11 methods. These include two traditional methods: CCA [2], JRL [23], and nine deep learning methods:DCCA [7], ACMR (2017) [13], CCL(2018) [25], FGCrossNet(2019) [8], CM-GANs(2019) [15], MHTN(2020) [30], DRSL(2021) [11], HCMSL(2021) [24] and DSCMR(2019) [14]. The detailed results are shown in Tables 1, 2 and 3. (The mAP scores obtained experimentally or provided by the authors of DSCMR [14] and DRSL [11]).

On Pascal Sentence Dataset, compared with the best mAP of the optimal DSCMR, DDAC improved by 2.9% on the Image2Text task and 1.3% on the Text2Image task, with an average improvement of 2.1%. On the Wikipedia Dataset, DDAC improved by 1.7% on the Image2Text task and 3.0% on the Text2Image task, with an average improvement of 2.35%, relative to the best mAP of the above optimal method DSCMR. On the NUS-WIDE-TC21 dataset, DDAC performs slightly better than the state-of-the-art DSCMR among other methods. In Table 4, it can be seen that the R@K (K=1, 3, 5) score of the DDAC method is better than the other methods in the three public datasets.The P-R curves also show that the DDAC method performs the best on all three datasets. See Figs. 2, 3 and 4 for details.

4.3 Impact of different components

In addition, we verify the effectiveness of each component in DDAC method through comparative experiments in the Pascal Sentence and Wikipedia datasets. In the experiment,some modules were removed to serve as a variant of DDAC:DDAC1 without projection loss L_d, DDAC2 without adversarial loss L_{adv}, DDAC3 and DDAC4 without the intra-modal discriminant loss L_{l1} and L_{l2} , respectively. DDAC5 does not have a rejector R, and L_d is to minimize the distance of all samples between modalities, that is to say, no label information is introduced. DDAC6 without inter-modal discrimination loss L_{imd}. DDAC7 replaces L_{imd} with J_2 from DSCMR [14]. Full-DDAC stands for the complete DDAC method.(Detailed test results are shown in Table 5)

It can be seen from the results that each module has a certain effect on the improvement of the mAP score on the Pascal Sentence and Wikipedia Datasets. Among them, the results obtained from DDAC1 and DDAC2 show that the projection loss has a slightly greater impact on DDAC than the adversarial loss.

In the label space, the effect of solving the intra-mode discrimination from the perspective of the probability distribution is better than that of minimizing the distance between the predicted label and the true label, which can be verified by the results of DDAC3 and DDAC4. Whether it is DDAC3 or DDAC4, its performance is significantly lower than the complete DDAC method, which can illustrate the effectiveness of the dual discrimination in this paper.

The experimental results of DDAC5 illustrate the effectiveness of introducing label information (including

Table 1 Comparison of the best mAP between the proposed method and the other 11 methods on Pascal Sentence Datasets

Method	Image2Text	Text2Image	Average
CCA [2]	0.225	0.227	0.226
JRL [23]	0.527	0.534	0.531
DCCA [7]	0.678	0.677	0.678
ACMR [13]	0.671	0.676	0.673
CCL [25]	0.576	0.561	0.569
CM-GANs [15]	0.603	0.604	0.604
FGCrossNet [8]	0.637	0.662	0.650
MHTN [32]	0.496	0.500	0.498
DRSL [11]	0.681	0.705	0.696
HCMSL [24]	0.699	0.712	0.710
DSCMR [14]	0.706	0.720	0.713
ourDDAC	0.735	0.733	0.734

Table 2 Comparison of the best mAP between the proposed method and the other 11 methods on Wikipedia Datasets

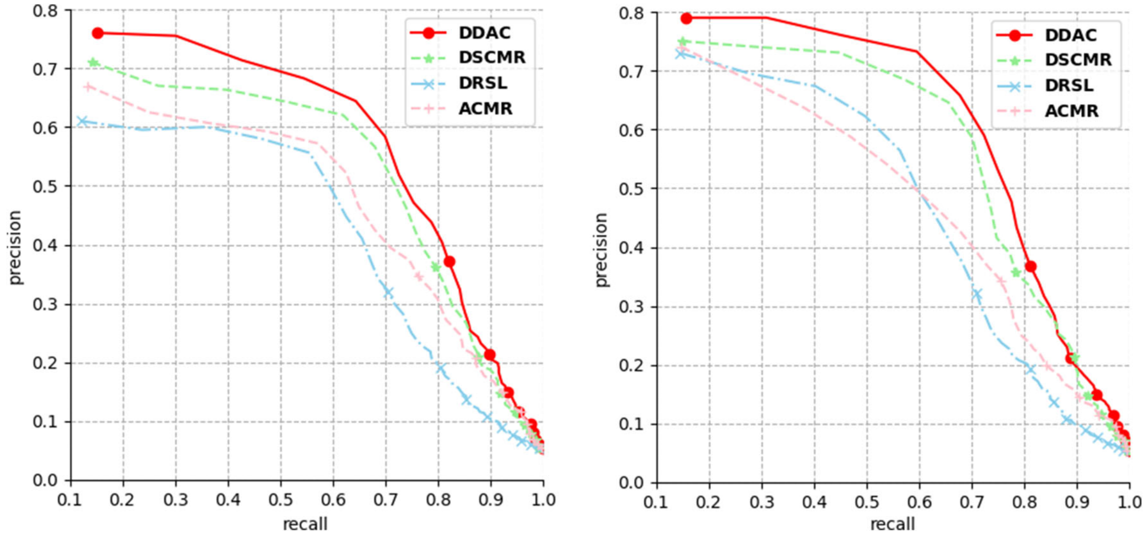
Method	Image2Text	Text2Image	Average
CCA [2]	0.134	0.133	0.134
JRL [23]	0.449	0.418	0.434
DCCA [7]	0.444	0.396	0.420
ACMR [13]	0.477	0.434	0.456
CCL [25]	0.504	0.457	0.481
CM-GANs [15]	0.521	0.466	0.494
FGCrossNet [8]	0.457	0.429	0.443
MHTN [32]	0.514	0.444	0.479
DRSL [11]	0.523	0.475	0.499
HCMSL [24]	0.524	0.476	0.500
DSCMR [14]	0.521	0.478	0.499
ourDDAC	0.538	0.508	0.523

Table 3 Comparison of the best mAP between the proposed method and the other 6 methods on NUS-WIDE-TC21 Datasets

Method	Image2Text	Text2Image	Average
CCA [2]	0.189	0.188	0.189
JRL [23]	0.429	0.376	0.401
DCCA [7]	0.448	0.465	0.457
ACMR [13]	0.544	0.538	0.541
DRSL [11]	0.540	0.552	0.546
DSCMR [14]	0.575	0.584	0.580
ourDDAC	0.582	0.592	0.587

Table 4 Comparing the retrieval results of R@K on the Wikipedia and pascal datasets

Dataset	Method	i2t-R@1	i2t-R@3	i2t-R@5	t2i-R@1	t2i-R@3	t2i-R@5
Pascal sentence	DDAC	0.75	0.83	0.89	0.80	0.91	0.95
	DSCMR	0.68	0.82	0.88	0.78	0.88	0.93
	DRSL	0.61	0.74	0.81	0.71	0.82	0.90
	ACMR	0.67	0.8	0.87	0.72	0.86	0.89
Wikipedia	DDAC	0.516	0.592	0.623	0.676	0.806	0.860
	DSCMR	0.500	0.554	0.588	0.707	0.783	0.872
	DRSL	0.493	0.586	0.606	0.595	0.788	0.863
	ACMR	0.462	0.591	0.615	0.616	0.785	0.837
NUS-WIDE-TC21	DDAC	0.696	0.804	0.842	0.650	0.742	0.769
	DSCMR	0.682	0.803	0.833	0.594	0.737	0.754
	DRSL	0.676	0.763	0.787	0.591	0.674	0.696
	ACMR	0.632	0.756	0.794	0.622	0.713	0.767



(a) Image2Text

(b) Text2Image

Fig. 2 P-R curves for Pascal Sentence Datasets

the rejector R) when reducing the heterogeneity difference between modalities.

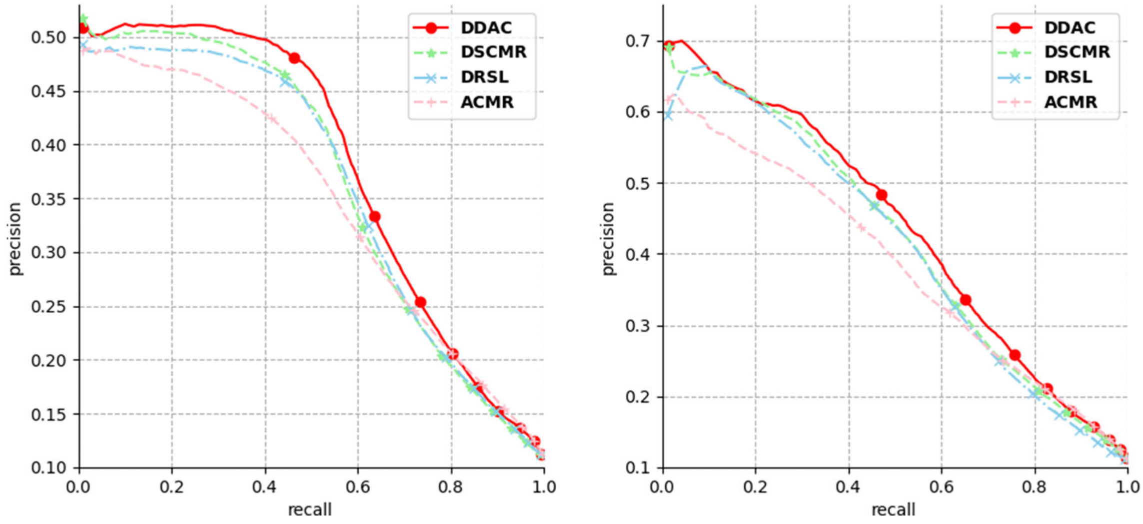
The results of DDAC6 show that the influence of inter-modal discriminant loss on the retrieval effect is relatively light, but it can not be ignored. As can be seen from the results of DDAC7, the effect of loss function J_2 in DSCMR is almost the same as that of L_{imd} in DDAC. However, L_{imd} is much simpler and can reduce the model's complexity slightly. J_2 is defined as:

$$J_2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(log\left(1+e^{\Gamma_{ij}}\right) - S_{ij}^{\alpha\beta}\Gamma_{ij}\right) \tag{10}$$

where, Γ_{ij} is used to calculate the cosine distance of samples between modalities. If the two samples have the same inter-modal category, the value of $S_{ij}^{\alpha\beta}$ is 1, otherwise 0.

4.4 Parameter analysis

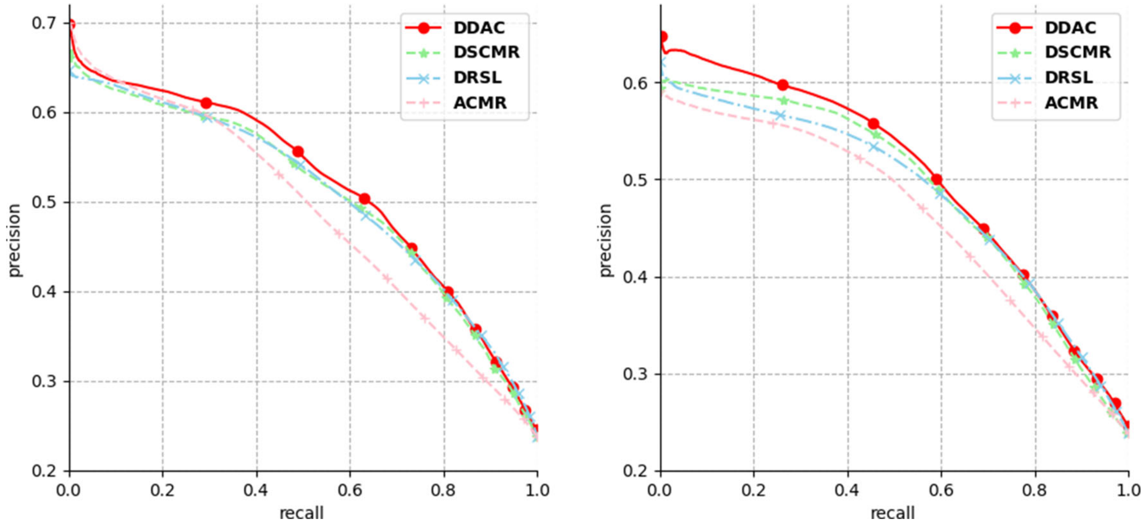
In the previous experiment, we set the parameters of the model in the objective function part. The parameters are divided into three groups. β and γ are used to control the contribution of projection loss and adversarial loss in the inter-modal consistency loss, parameters λ and η control the contribution of two label losses in the intra-modal



(a) Image2Text

(b) Text2Image

Fig. 3 P-R curves for Wikipedia Datasets



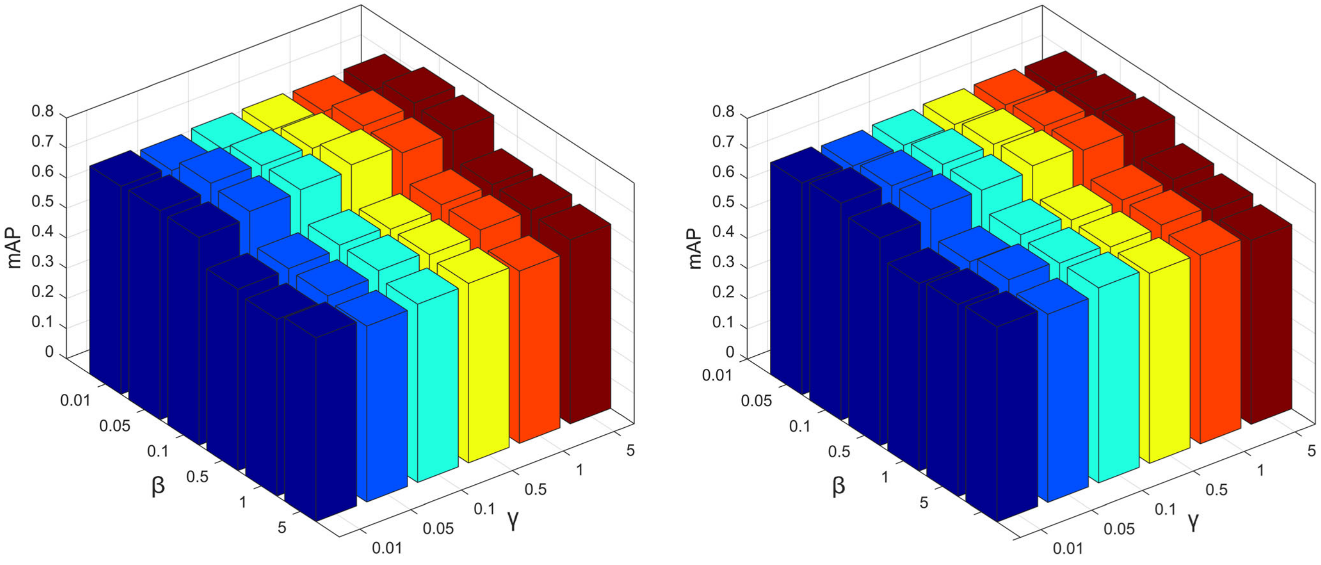
(a) Image2Text

(b) Text2Image

Fig. 4 P-R curves for NUS-WIDE-TC21 Datasets

Table 5 The MAP score in the DDAC comparison experiment

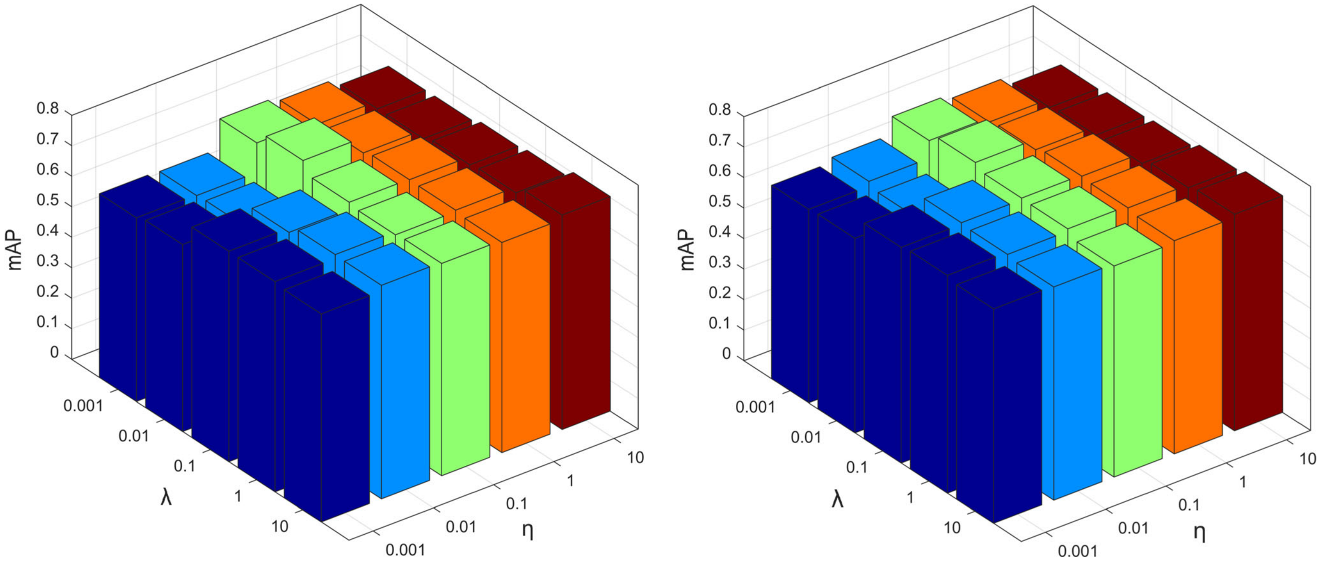
Method	Pascal		Wikipedia	
	Image2Text	Text2Image	Image2Text	Text2Image
DDCA1	0.691	0.706	0.515	0.477
DDCA2	0.704	0.716	0.524	0.492
DDCA3	0.709	0.716	0.522	0.491
DDCA4	0.614	0.640	0.523	0.488
DDCA5	0.726	0.725	0.530	0.497
DDCA6	0.713	0.719	0.522	0.493
DDCA7	0.734	0.733	0.536	0.507
Full-DDAC	0.735	0.733	0.538	0.508



(a) Image2Text

(b) Text2Image

Fig. 5 mAP changes when β and γ are different



(a) Image2Text

(b) Text2Image

Fig. 6 mAP changes when λ and η are different

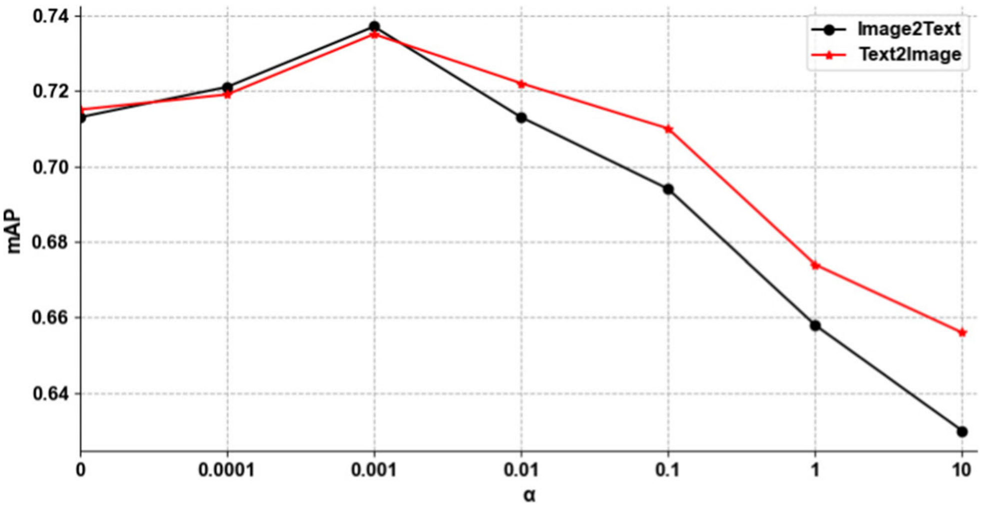
discriminant loss, and α control the contribution of inter-modal discriminant loss to the overall objective function. We selected Pascal Sentence datasets as experimental data and analyzed the influence of these parameters on the model results in the training process.

We divide the parameters into three groups according to what the objective function does, which are set as:$\lambda, \eta \in \{0.001, 0.01, 0.1, 1, 10\}$, $\beta, \gamma \in \{0.001, 0.05, 0.1, 0.5, 1, 5\}$, $\alpha \in \{0.001, 0.05, 0.1, 0.5, 1, 5\}$.

The evaluation method for the first two groups of parameters is to first fix the other two groups of parameters, then fix a parameter of the group, and then change the size of the other parameter. For the third group of parameters, fix the other two groups of parameters and change the value of α. Figures 5 to 7 show the corresponding results.

As can be seen from Figs. 5, 6, and 7, for the first group of parameters,β can get a good mAP score when it is 0.05 and 0.1, while γ has no significant influence on the MAP score. After repeated experiments, we finally determined that $\beta = 0.1, \gamma = 0.5$;In the experimental results shown in Fig. 3, a better mAP canbe obtained when the second set of parameters $\lambda = 0.01$ and $\eta = 0.1$ are used; It can be clearly seen from Fig. 4 that, for the third group of parametersα, when, $\alpha = 0.001$ the best mAP can be obtained.

Fig. 7 mAP changes when α are different



5 Conclusion

In this paper, we presents a cross-modal retrieval method based on adversarial learning with dual discrimination. DDAC minimizes the projection loss and eliminates the inter-modal heterogeneity through adversarial learning, while the additional rejector prevents the bridging of the inter-modal heterogeneity between the projections of different semantic features in the adversarial learning process. Then, we minimize the projection distances of features with the same semantic labels between modalities by cosine distance and maximize the projection distances of different semantic features. Finally, in the label space, DDAC double discriminates the semantic labels of feature projection items from the perspectives of probability distribution and distance respectively, which can effectively ensure the classification information in the modality is unchanged. Finally, the proposed method is tested on two widely used datasets to verify the superiority of the DDAC method in cross-modal retrieval performance.

References

1. Wang K, Yin Q, Wang W, Wu S, Wang L (2016) A Comprehensive Survey on Cross-modal Retrieval. arXiv:1607.06215
2. Hardoon D, Szedmák S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Comput 16:2639–2664
3. Chua T, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: A real-world web image database from National University of Singapore. CIVR '09
4. Chen W, Liu Y, Bakker EM, Lew MS (2021) Integrating Information Theory and Adversarial Learning for Cross-modal Retrieval. arXiv:2104.04991
5. Zhang X, Lai H, Feng J (2018) Attention-Aware Deep Adversarial Hashing for Cross-Modal Retrieval. ECCV
6. Zhang Y, Feng Y, Liu D, Shang J, Qiang B (2020) FRWCAE: Joint faster-RCNN and Wasserstein convolutional auto-encoder for instance retrieval. Appl Intell 50:2208–2221
7. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep Canonical Correlation Analysis. ICML
8. He X, Peng Y, Xi-e L (2019) A new benchmark and approach for fine-grained cross-media retrieval. In: Proceedings of the 27th ACM international conference on multimedia
9. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-Modal Retrieval with CNN visual features: A new baseline. IEEE Trans Cybern 47:449–460
10. Wang C, Yang H, Meinel C (2015) Deep semantic mapping for cross-modal retrieval. In: 2015 IEEE 27th international conference on tools with artificial intelligence (ICTAI), pp 234–241
11. Wang X, Hu P, Zhen L, Peng D (2021) DRSL: Deep relational similarity learning for cross-modal retrieval. Inf Sci 546:298–311
12. Castellano G, Fanelli A, Sforza G, Torsello MA (2015) Shape annotation for intelligent image retrieval. Appl Intell 44:179–195
13. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT (2017) Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on multimedia
14. Zhen L, Hu P, Wang X, Peng D (2019) Deep supervised cross-modal retrieval. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10386–10395
15. Peng Y, Qi J, Yuan Y (2019) CM-GANS. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15:1–24
16. Li C, Deng C, Li N, Liu W, Gao X, Tao D (2018) Self-supervised adversarial hashing networks for cross-modal retrieval. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 4242–4251
17. Bai C, Zeng C, Ma Q, Zhang J, Chen S (2020) Deep Adversarial discrete hashing for Cross-Modal retrieval. In: Proceedings of the 2020 international conference on multimedia retrieval
18. Simonyan K, Zisserman A (2015) Very deep convolutional networks for Large-Scale image recognition. CoRR, arXiv:1409.1556
19. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed Representations of Words and Phrases and their Compositionality. NIPS
20. Kang P, Lin Z, Yang Z, Fang X, Bronstein A, Li Q, Liu W (2021) Intra-class low-rank regularization for supervised and semi-supervised cross-modal retrieval. Appl Intell, pp 1–22
21. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. CoRR, arXiv:1412.6980
22. Rashtchian C, Young P, Hodosh M, Hockenmaier J (2010) Collecting Image Annotations Using Amazon's Mechanical Turk. Mturk@HLT-NAACL
23. Zhai X, Peng Y, Xiao J (2014) Learning Cross-Media joint representation with sparse and semisupervised regularization. IEEE Trans Circuits Syst Video Technol 24:965–978
24. Zhang C, Song J, Zhu X, Zhu L, Zhang S (2021) HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval. ACM Trans Multimedia Comput Commun Appl (TOMM) 17:1–22
25. Peng Y, Qi J, Huang X, Yuan Y (2018) CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. IEEE Trans Multimedia 20:405–420
26. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-Modal Retrieval with CNN visual features: A new baseline. IEEE Trans Cybern 47:449–460
27. Wang W, Yang X, Ooi B, Zhang D, Zhuang Y (2015) Effective deep learning-based multi-modal retrieval. The VLDB J 25:79–101
28. Li Z, Lu W, Bao E, Xing W (2015) Learning a Semantic Space by Deep Network for Cross-media Retrieval. DMS
29. Pereira JC, Coviello E, Doyle G, Rasiwasia N, Lanckriet G, Levy R, Vasconcelos N (2014) On the role of correlation and abstraction in Cross-Modal multimedia retrieval. IEEE Trans Pattern Anal Mach Intell 36:521–535
30. Huang X, Peng Y, Yuan M (2020) MHTN: Modal-Adversarial Hybrid transfer network for Cross-Modal retrieval. IEEE Trans Cybern 50:1047–1059
31. Zhou Y, Feng Y, Zhou M, Qiang B, UL, Zhu J (2021) Deep adversarial quantization network for Cross-Modal retrieval. In: ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4325–4329
32. Zhang JG, Peng Y, Yuan M (2020) SCH-GAN: Semi-Supervised Cross-Modal Hashing by generative adversarial network. IEEE Trans Cybern 50:489–502
33. Song G, Wang D, Tan X (2019) Deep memory network for Cross-Modal retrieval. IEEE Transactions on Multimedia 21:1261–1275
34. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-Modal Retrieval with CNN visual features: a new baseline. IEEE Trans Cybern 47:449–460