mathematics

MDPI

Article

Deep Adversarial Learning Triplet Similarity Preserving Cross-Modal Retrieval Algorithm

Guokun Li [1], Zhen Wang [1,2,*], Shibo Xu [1], Chuang Feng [1], Xiaohan Yang [1], Nannan Wu [1] and Fuzhen Sun [1]

[1] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China; 20505010623@stumail.sdut.edu.cn (G.L.); 20110507056@stumail.sdut.edu.cn (S.X.); 20505010621@stumail.sdut.edu.cn (C.F.); 21505020639@stumail.sdut.edu.cn (X.Y.); 21505020635@stumail.sdut.edu.cn (N.W.); sunfuzhen@sdut.edu.cn (F.S.)

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

* Correspondence: wzh@sdut.edu.cn

Abstract: The cross-modal retrieval task can return different modal nearest neighbors, such as image or text. However, inconsistent distribution and diverse representation make it hard to directly measure the similarity relationship between different modal samples, which causes a heterogeneity gap. To bridge the above-mentioned gap, we propose the deep adversarial learning triplet similarity preserving cross-modal retrieval algorithm to map different modal samples into the common space, allowing their feature representation to preserve both the original inter- and intra-modal semantic similarity relationship. During the training process, we employ GANs, which has advantages in modeling data distribution and learning discriminative representation, in order to learn different modal features. As a result, it can align different modal feature distributions. Generally, many cross-modal retrieval algorithms only preserve the inter-modal similarity relationship, which makes the nearest neighbor retrieval results vulnerable to noise. In contrast, we establish the triplet similarity preserving function to simultaneously preserve the inter- and intra-modal similarity relationship in the common space and in each modal space, respectively. Thus, the proposed algorithm has a strong robustness to noise. In each modal space, to ensure that the generated features have the same semantic information as the sample labels, we establish a linear classifier and require that the generated features' classification results be consistent with the sample labels. We conducted cross-modal retrieval comparative experiments on two widely used benchmark datasets—Pascal Sentence and Wikipedia. For the image to text task, our proposed method improved the mAP values by 1% and 0.7% on the Pascal sentence and Wikipedia datasets, respectively. Correspondingly, the proposed method separately improved the mAP values of the text to image performance by 0.6% and 0.8% on the Pascal sentence and Wikipedia datasets, respectively. The experimental results show that the proposed algorithm is better than the other state-of-the-art methods.

Keywords: cross-modal retrieval; generative adversarial network; triplet similarity preserving; deep representation learning

MSC: 68T45

Citation: Li, G.; Wang, Z.; Xu, S.; Feng, C.; Yang, X.; Wu, N.; Sun, F. Deep Adversarial Learning Triplet Similarity Preserving Cross-Modal Retrieval Algorithm. *Mathematics* **2022**, *10*, 2585. https://doi.org/10.3390/math10152585

Academic Editors: Xinchao Zhao, Xingquan Zuo, Yinan Guo and Kunpeng Kang

Received: 7 June 2022
Accepted: 22 July 2022
Published: 25 July 2022

1. Introduction

Multimedia data with different modalities, such as image, text, video, and audio, are mixed together and represent comprehensive knowledge needed in order to perceive the real world [1–6]. Exploring the cross-modal retrieval between image and natural language has recently attracted great interest among researchers, due to its great importance in various applications, such as bi-directional image and text retrieval, natural language object retrieval, image captioning, and visual question answering [7]. The image–text cross-modal retrieval algorithms can return samples with the same semantic label, but with a different

modality from the query sample. Figure 1 illustrates the image–text cross-modal retrieval tasks, which include image retrieving text and text retrieving image. In Figure 1, there are airplane images as well as texts describing the airplane number, flight status, airplane type, etc. A critical task for cross-modal retrieval is to measure the similarity between the image and text. To achieve this goal, many existing cross-modal retrieval algorithms propose to deep-learning network that can map different modal features into the common space. As shown in Figure 1, we mapped airplane images and texts into the common space using the pre-trained deep network, and measured their similarity relationship by computing the distances among the different modal features.

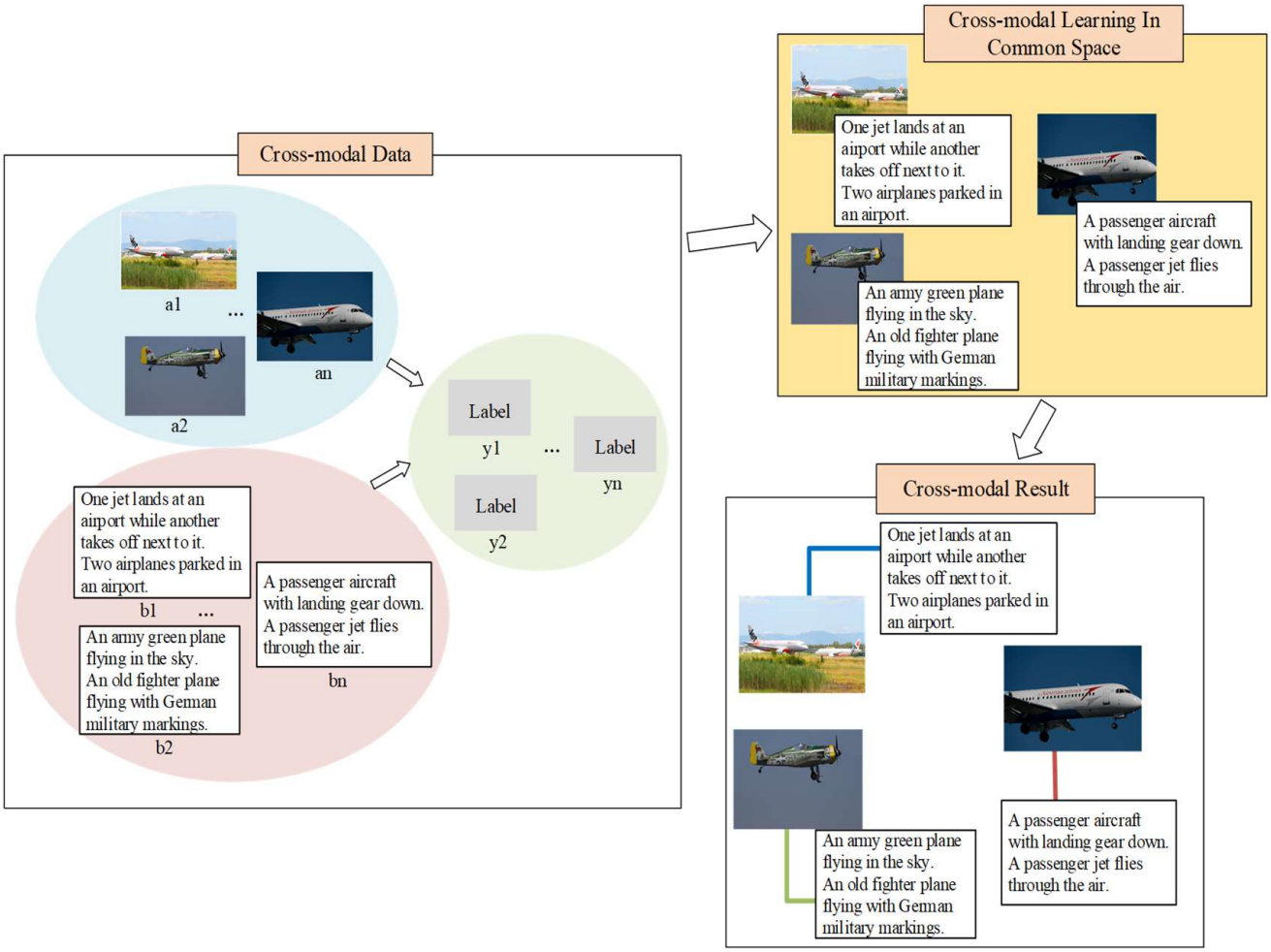


Figure 1. The cross-modal retrieval task includes image retrieving text and text retrieving images.

In order to obtain an excellent cross-modal retrieval performance, different modal features need to be mapped into the common space by preserving their original semantic neighbor information. Firstly, the cross-modal retrieval algorithm aims to retrieve the nearest neighbor with a different modality from the query sample. So, the semantic relationship among the different modal samples needs to be preserved. Secondly, the cross-modal retrieval algorithms usually return more than one neighbor. Thus, the preservation of the intra-modal semantic relationship also needs to be taken into consideration.

Generally, most cross-modal retrieval algorithms only focus on preserving different modal similarity relationships in the common space, while ignoring the similarity relationship preserving problem in each single-modal space. As a result, many dissimilar samples could have similar feature representations. Unfortunately, the noise would further make these dissimilar samples have almost the same feature representations, and these dissimilar samples would be incorrectly returned as the nearest neighbors. Furthermore, different

modal features have inconsistent distributions and representations. However, many existing methods learn the common representations without aligning their distributions [8]. To solve the above-mentioned problems, we propose a novel cross-modal network based on an adversarial learning algorithm, as shown in Figure 2.

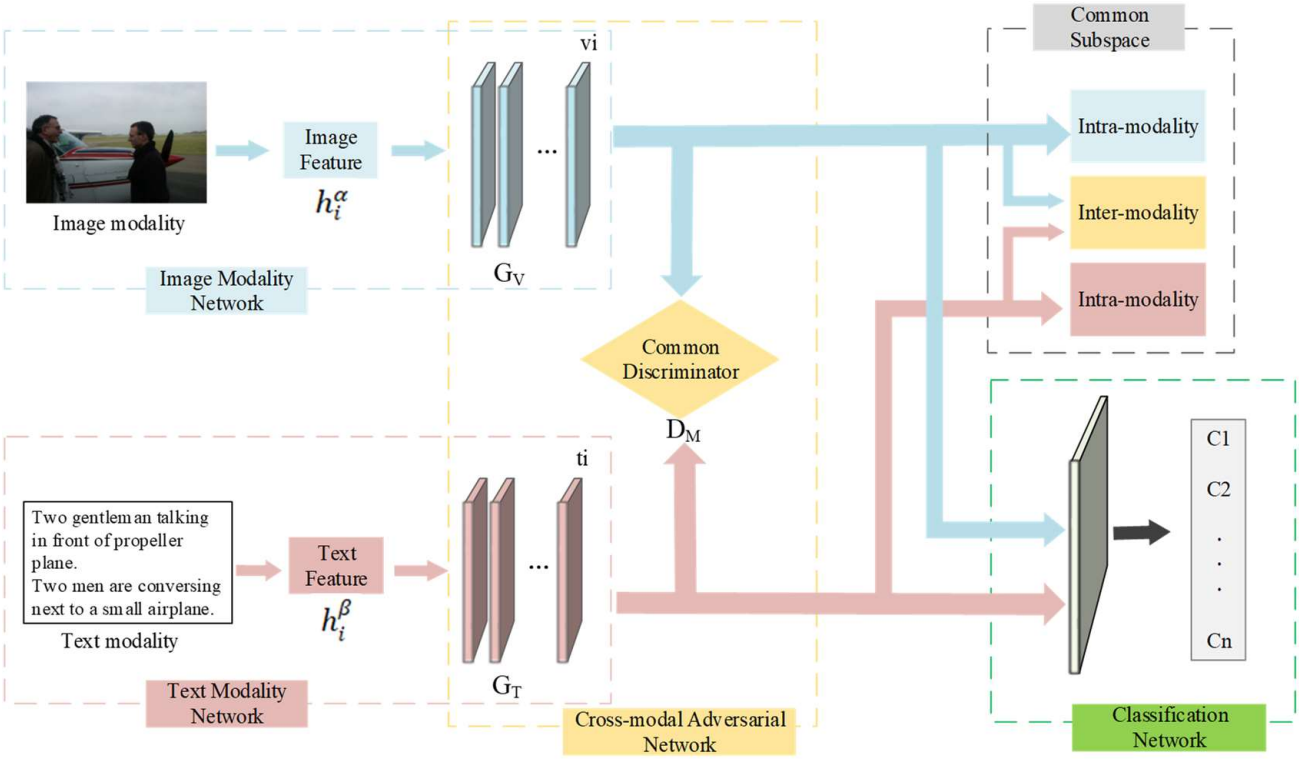


Figure 2. The framework of the proposed cross-modal retrieval algorithm.

The proposed method consists of the following parts: the image modal network, the text modal network, the cross-modal adversarial network, the common space network, and the classification network. (1) For the image modality, we pre-trained VGG-19 [9] on ImageNet and considered the 4096-dimensional vector generated by the fc7 layer as the image modal feature $h_i{}^{\alpha}$. Then, three full-connected layers were employed to learn the image representation v_i in the common space. (2) For the text modality, we used the Doc2Vec [10] model to generate the 300-dimensional text modal feature $h_i{}^{\beta}$. Similarly, three full-connected layers were employed to learn the text representation t_i in the common space. (3) For the cross-modal adversarial network, we considered the fully connected layers in each modality as the image generator G_v and the text generator G_t. Furthermore, we established a common discriminator D_M to distinguish the input feature's modality information. Adversarial learning utilized the mini-max mechanism to train the generator networks G_v and G_t and the discriminator network D_M. During the iterative training process, the generators aimed to minimize the probability of being correctly recognized by the discriminator. In contrast, the discriminator tried to maximize the probability of correctly recognizing the sample's modal information. When the algorithm converged, we could align the feature distributions between the different modalities. (4) For the common space network, we established the triplet similarity preserving function to preserve the inter- and intra-modal similarity relationship. (5) For the classification network, we utilized a linear projection function to classify the sample features, and required the classification results to be consistent with the original semantic labels. Thus, the generated features had the same semantic information as the sample's label.

The main contributions were as follows:

(1) We employed the generative adversarial network to learn different modal features in the common space, which could reduce the distribution difference between the different modal features using the mini-max mechanism;
(2) In the common, image, and text spaces, we separately designed the triplet similarity preserving function to preserve both the inter- and intra-modal similarity relationship. This could also boost the retrieval results for robustness to noise;
(3) To avoid the loss of semantic information during learning image and text features, we established a linear function to predict the generated feature labels, and required that the prediction labels were identical to the original semantic labels.

2. Related Work

The cross-modal retrieval algorithm can retrieve similar samples with different modalities, which helps to comprehensively perceive and recognize the query sample. However, different modal features have different distributions and diverse representations, which lead to the heterogeneity gap. Thus, to directly compute the different modal similarity relationship, different modalities need to be mapped into a common space [7,11–13]. Traditionally, linear projection functions with an optimized target statistical value are utilized to map different modal features into a common space [14]. The canonical correlation analysis (CCA) [11] finds a linear combination to maximize the pairwise correlations between the two data sets, and associates different modal features by projecting them into a common space. The cross-modal factor analysis (CFA) [12] learns different modalities' common space by minimizing the data pair's Frobenius norm. Joint representation learning (JRL) [13] learns the sparse projection matrices, and adds the unlabeled data to improve the diversity of the training data. The deep-learning-based cross-modal retrieval methods employ the scalable nonlinear transformation to learn the sample's content representation [14]. Ngiam et al. [15] proposed a bimodal auto-encoder to learn different modal correlations, and applied the restricted Boltzmann machine (RBM) to generate the common space. The multimodal deep neural network (MDNN) [16] utilizes the deep convolutional neural network (CNN) to learn the image feature and employs the neural language model (NLM) to learn the text feature. Furthermore, MDNN establishes the correlation between different modal features by projecting them into a common space.

Generally, the cross-modal retrieval algorithms can be divided into three categories, namely unsupervised approaches [11,17,18], pairwise approaches [19,20], and supervised approaches [14,21]. The unsupervised methods directly exploit different modal feature information to learn their common representations [7]. For example, CCA [11], Deep-CCA [17], and Deep Canonical Correlated Auto-encoder (DCCAE) [18] utilize the correlations between heterogeneous data to learn the common representations. The pairwise-based methods, such as the multiview metric learning with global consistency and local smoothness (MVML-GL) method [22] and the modality-specific deep structure (MSDS) method [20], generate the similarity metrics according to the similarity relationship between different modal sample pairs [7]. The supervised methods try to preserve the original semantic label information in the common space. Sharma et al. [23] proposed the generalized multi-view analysis (GMA) method based on CCA, which supervises learning the common representations using the semantic category labels. In [14] and [21], generative adversarial networks [24] are used to generate different modal features and reduce the distribution difference.

3. The Proposed Method

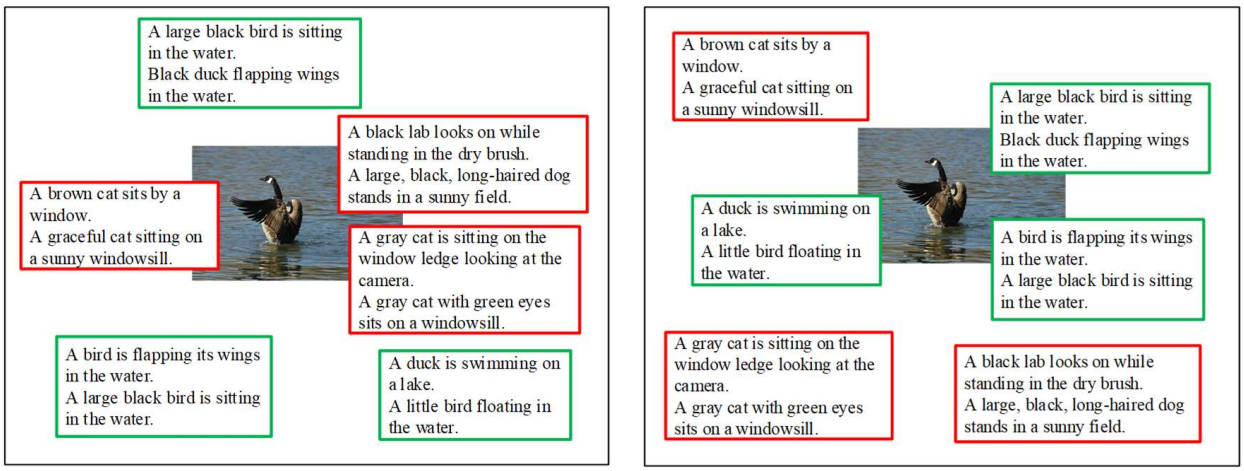
3.1. Notations

The notations used in this paper are given as follows. $O = \{a_i, b_i\}_{i=1}^{n}$ denotes n pairs of image and text. a_i is the image and b_i is the text. $y_i = [y_{i1}, y_{i2}, \ldots, y_{ic}] \in \mathbb{R}^c$ is the semantic label vector, where c is the number of the categories. If the *i-th* instance belongs to the *j-th* category, $y_{ij} = 1$, otherwise $y_{ij} = 0$. $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{c \times n}$ is the label matrix. $f(x)$ and

$g(x)$ represent the image and text feature learning network. $v_i = f(ai) \in \mathbb{R}^{dv}$ is the feature of image a_i and $V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{dv \times n}$ is the image feature matrix. $t_i = g(bi) \in \mathbb{R}^{dt}$ is text feature and $T = [t_1, t_2, \ldots, t_n] \in \mathbb{R}^{dt \times n}$ is the text feature matrix. d_v and d_t represent the number of dimensions.

3.2. The Triplet Similarity Relationship Preserving Function

The cross-modal retrieval algorithm measures the similarity relationship between different modal samples in the common space, and returns the samples with minimal distance between the nearest neighbors. To obtain an excellent cross-modal retrieval performance, the distance between the same category samples should be smaller than that between different categories samples. In Figure 3a, the anchor and positive samples belong to the same category. In contrast, the anchor and negative samples belong to different categories. Thus, in Figure 3b, the anchor's feature should be similar to the positive sample and different from the negative sample. This ensures the positive samples can first be returned as the nearest neighbors of the anchor.



(**a**) (**b**)

Figure 3. The distance relation preserving among triplet samples in the common space. (**a**) The sample distribution without the triplet similarity preserving constraint. (**b**) The sample distribution with the triplet similarity preserving constraint.

To achieve the above goal, we proposed to simultaneously preserve the inter- and intra-modal triplet similarity relationship.

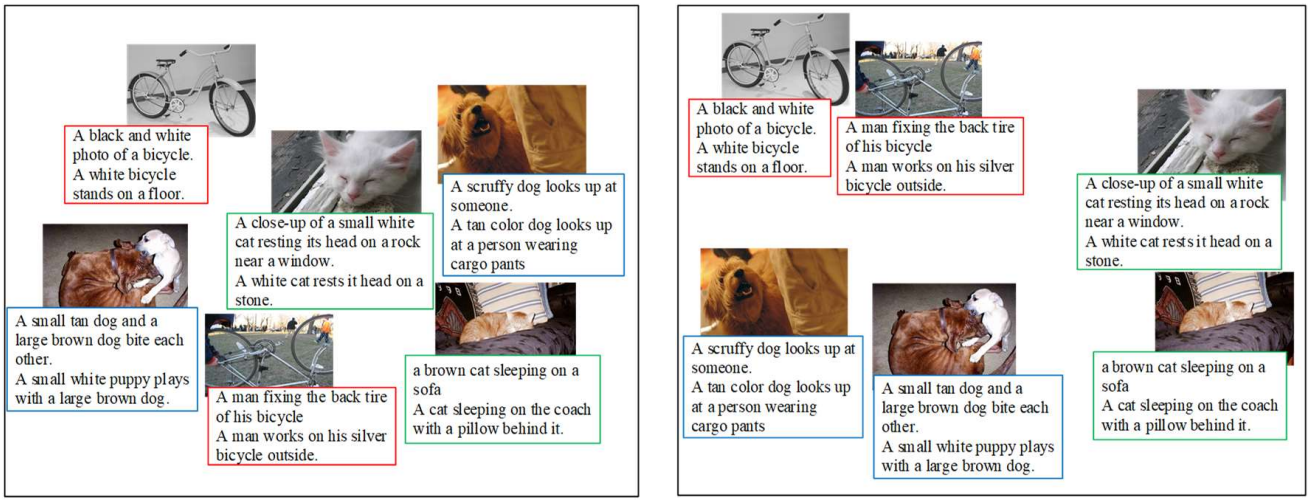
In this paper, we designed the triplet loss function between the image–text, image−image, and text−text, respectively.

The triplet similarity preserving function between image–text is defined as in Equation (1):

$$L_O = \sum_V [d(V_a, T_p) - d(V_a, T_n) + \alpha]_+ + \sum_T [d(T_a, V_p) - d(T_a, V_n) + \alpha]_+ \quad (1)$$

In Equation (1), L_o represents the image–text loss. $d(\cdot)$ is the Euclidean distance. V_a, V_p, and V_n represent the anchor, positive, and negative images, respectively. T_a, T_p, and T_n represent the anchor, positive, and negative texts, respectively. α is the error margin.

Generally, most of the existing methods only focus on preserving the similarity relationship between different modal samples, while not preserving the similarity relationship among the same modal samples [25]. As shown in Figure 4a, samples that belong to different categories may have a small distance, which may lead to incorrect retrieval results.



(**a**) (**b**)

Figure 4. Similarity relationship preserving among samples. (**a**) The sample distribution with only inter-modal similarity preserving constraint. (**b**) The sample distribution with both inter- and intra-modal similarity preserving constraints.

To solve the above problem, we designed the intra-modal triplet similarity preserving function as in Equation (2). It ensures that the distances among similar samples are smaller than those among dissimilar samples in the image- or text- modality, respectively.

$$\begin{aligned} L_v &= \sum_V \left[d\left(V_a, V_p\right) - d(V_a, V_n) + \alpha \right]_+ \\ L_t &= \sum_T \left[d\left(T_a, T_p\right) - d(T_a, T_n) + \alpha \right]_+ \end{aligned} \tag{2}$$

In Equation (2), L_v is the image-modal triplet similarity preserving function, and L_t is the text-modal triplet similarity preserving function.

In this paper, we aimed to preserve both the inter- and intra-modal triplet similarity relationship and to define the triplet similarity preserving objective function as in Equation (3).

$$L_{\mathrm{Re}t} = L_o + L_v + L_t \tag{3}$$

By simultaneously minimizing the value of L_o, L_v, and L_t, we ensured the distance between the same category samples is small and between the different categories samples it is large, as shown in Figure 4b. As a result, the proposed method is robust to noise.

3.3. The Minimal Semantic Information Loss

The proposed method utilizes the floating-point feature representing the sample content in the triplet similarity preserving function. Therefore, to guarantee the cross-modal search performance, the sample features should have the same semantic information as its labels.

In this paper, we learned the samples' features using a deep neural network. Due to information loss, the deep learning features may not accurately preserve the original semantic information.

To solve the above problem, we employed a linear projection function to classify the deep feature, and required that the predicted label be identical to the sample label, as

shown in Figure 5. We formulate the above procedure as in Equation (4), which minimizes the difference between the features' classification results and the samples' semantic labels.

$$L_{Dis} = \frac{1}{n}\|P^T V - Y\|_F + \frac{1}{n}\|P^T T - Y\|_F \quad (4)$$

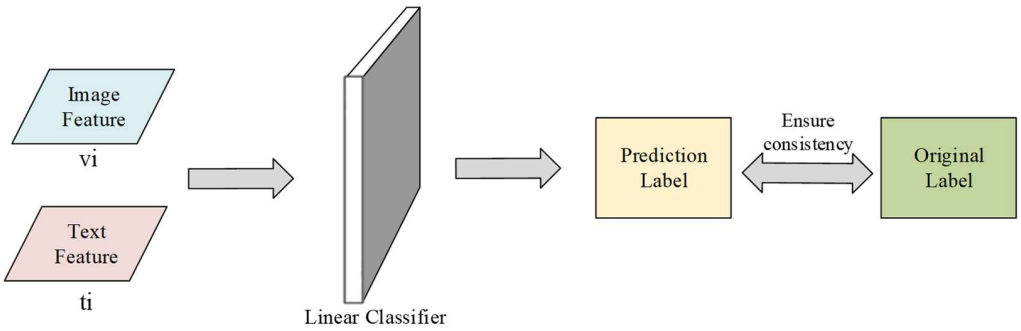


Figure 5. The framework for preserving the semantic information between the generated features and different modal samples.

In Equation (4), $\|\cdot\|_F$ is the Frobenius norm. P is the matrix of the linear projection function.

3.4. The Cross-Modal Adversarial Learning

To directly compute the similarity relationship between the different modal samples, different modal features should have the same distribution in the common space.

As generative adversarial networks (GANs) have a strong ability for modeling data distribution and learning discriminative representation [26], we used the GANs to align the distribution between the different modal features. Figure 6 shows the adversarial learning procedure. We regarded the fully connected layers of the image modal network as the image generator G_V and the fully connected layers of the text modal network as the text generator G_T. Initially, the discriminator D_M regards the image features generated by G_V as the real samples, and considered the text features generated by G_T as the fake samples. In Equation (5), adversarial loss is defined as the difference between different modal feature distributions. Both the generator and discriminator utilize Equation (5) as the objective function. During the training procedure, we employed the min−max mechanism. The discriminator tries to maximize the objective value and the generators aims to minimize the objective value. When the algorithm converges, D_M can only randomly distinguish the sample's modal information. Finally, the image and text features have the same distribution.

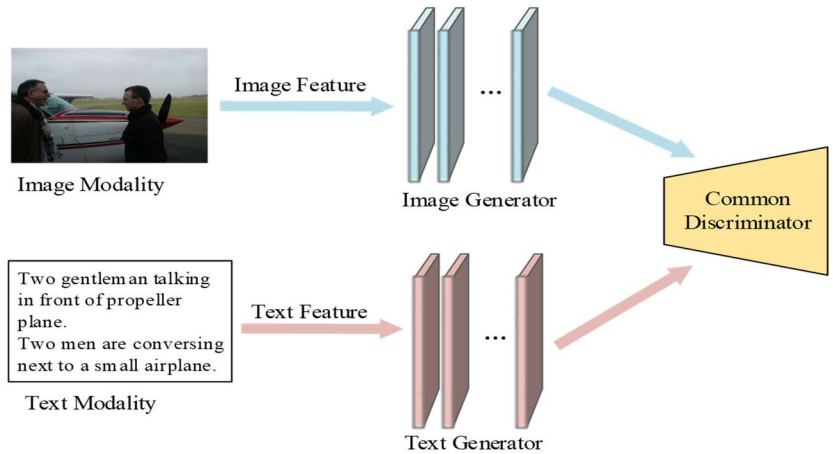


Figure 6. The adversarial learning network. It reduces the difference between the different modal feature distributions using the min−max mechanism.

$$\begin{aligned} L_{Adv} &= E_{v\sim P_{image}}[\log D_M(G_V(v))] \\ &+E_{t\sim P_{text}}[\log(1 - D_M(G_T(t)))] \end{aligned} \tag{5}$$

In this paper, we define the final objective function L as in Equation (6), which can achieve the following three tasks: (1) preserve the inter- and intra-modal relative similarity relationship, (2) minimize the semantic loss during learning the deep features, and (3) align the different modal features' distribution.

$$L = L_{Dis} + \lambda L_{\mathrm{Ret}} + \eta L_{Adv} \tag{6}$$

In Equation (6), λ and η are the weight parameters. We optimized the objective function L using the stochastic gradient descent algorithm [27].

4. The Comparative Experiments

In this paper, we conducted comparative experiments on widely used datasets, namely the Pascal Sentence dataset [28] and the Wikipedia dataset [29]. These two datasets are publicly available. The retrieval tasks included image retrieving text and text retrieving image. To verify the effectiveness of our proposed methods, we employed five state-of-the-art methods, namely, CCA [11], JRL [13], CMDN [30], Deep-SM [19], and DSCMR [7], as the comparative methods. We implemented the model development and data analysis using the PyTorch deep learning framework.

4.1. The Datasets and Settings

The Pascal Sentence dataset [28] includes 1000 image–text pairs and a total of 20 categories. Each image is described by five sentences in a document. We divided the Pascal Sentence dataset into three parts, namely, the training, test, and validation sets. We randomly selected 800 image–text pairs as the training set and 100 image–text pairs as the test set. The proposed method generated the 4096-dimensional vector as the image feature, and the text feature had 300-dimensions.

The Wikipedia dataset [29] includes 2866 image–text pairs that can be divided into 10 categories. Each pair consists of an image and several text paragraphs. We randomly selected 2173 pairs as the training set, and considered the remaining 693 pairs as the test set. The dimension of the image feature was 4096, and the text feature had 300-dimensions.

The proposed algorithm employed the deep adversarial network to learn different modal features in the common space. Both G_V and G_T had two fully connected layers, and utilized $tanh(\cdot)$ as the activate function. At the end, G_V and G_T had a fully connected layer that shared the weight values. The discriminator D_M consisted of three fully connected layers and employed the sigmoid function at the activation layer. For the triplet similarity preserving function in the common space, the value of α was set as 0.3. For the objective function L, $\lambda = 0.001$, $\eta = 0.1$.

4.2. Evaluation Metric

In this paper, we used mAP (mean average precision) and PR (precision−recall) curves to measure the cross-modal retrieval performance.

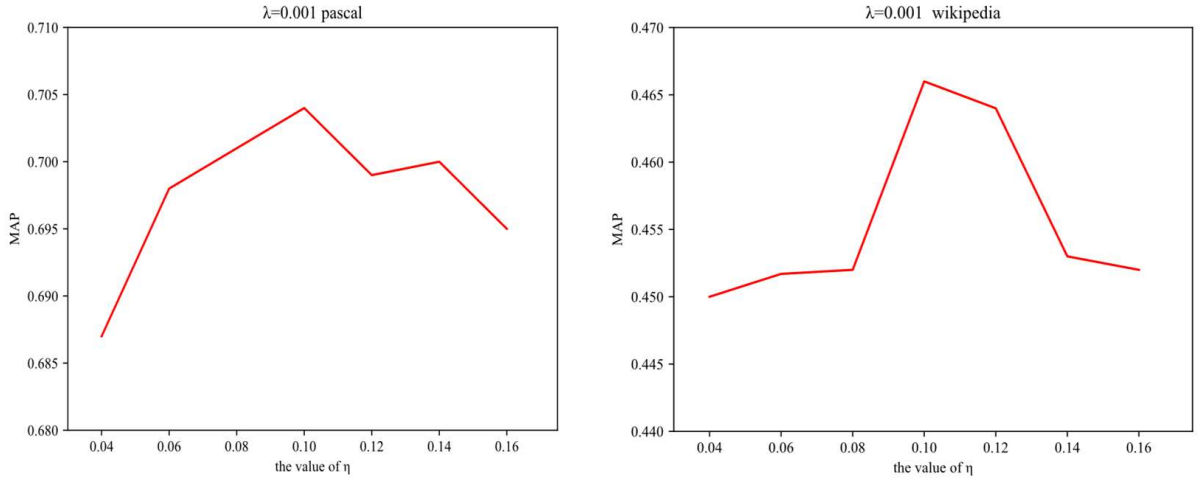
MAP [31] is the mean value of the average precision of all of the query sample retrieval results, and its definition is shown in Equation (7).

$$\mathrm{m}AP = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{K_i}\sum_{j=1}^{K_i}\frac{j}{rank(j)} \tag{7}$$

Q represents the number of query samples. K_i represents the number of the *i-th* query sample's ground truth. j is the numerical order of the *j-th* ground truth. *rank (j)* returns the ranking order of the *j-th* true positive data in the retrieval result.

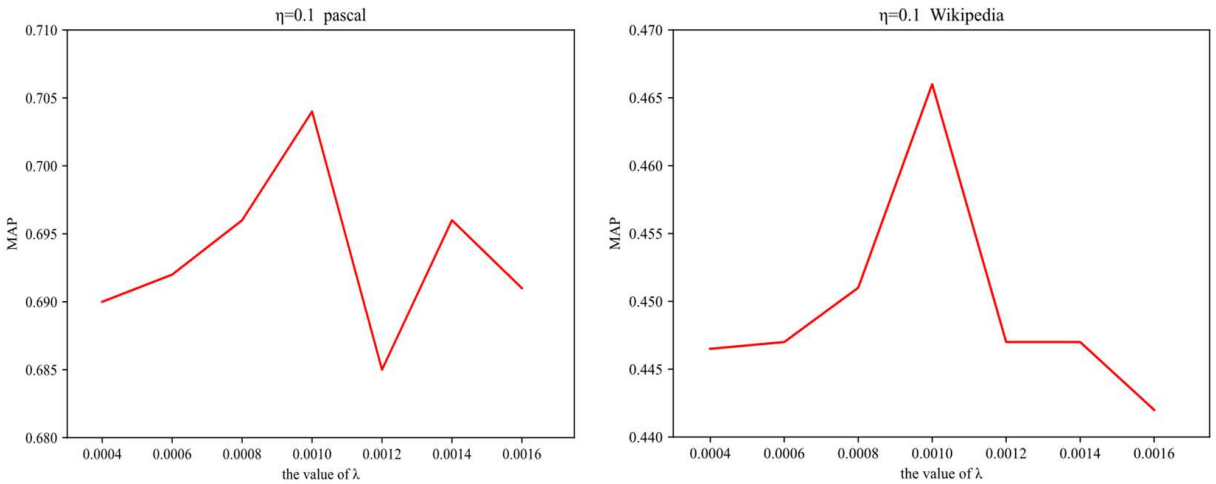
4.3. The Parameters Values

In this paper, we set the parameters λ and η to balance the effect between the triplet similarity preserving the loss and adversarial loss during the training process. We compared the cross-modal retrieval performances with different values of λ and η on both the Pascal Sentence and Wikipedia datasets. The experimental results are shown in Figures 7 and 8. In Figure 7, the value of λ is fixed and the value of η gradually increases from 0.04 to 0.16. The best cross-modal retrieval performance occurs when η = 0.1. In Figure 8, η is fixed and the value of λ changes from 0.0004 to 0.0016. When λ = 0.001, we achieved the best cross-modal retrieval performance. As described above, we set the value of η as 0.1 and λ as 0.001 in this paper.



(**a**) (**b**)

Figure 7. The cross-modal retrieval performances with different η values. (**a**) The cross-modal retrieval performances with different η values and λ = 0.001 on the Pascal Sentence dataset. (**b**) The cross-modal retrieval performances with different η values and λ = 0.001 on the Wikipedia dataset.



(**a**) (**b**)

Figure 8. The cross-modal retrieval performances with different λ values. (**a**) The cross-modal retrieval performances with different λ values and η = 0.1 on the Pascal Sentence dataset. (**b**) The cross-modal retrieval performances with different λ values and η = 0.1 on the Wikipedia dataset.

4.4. Experimental Results and Analysis

The comparison algorithms include CCA [11], JRL [13], CMDN [30], Deep-SM [19] and DSCMR [7]. CCA [11] learns the common space which maximizes the pairwise interrelationships between two sets of heterogeneous data. JRL [13] learns different modal features by multi-metric learning. CCA and JRL belong to the traditional methods. CMDN [30] employs both the intra- and inter-modal information, and utilizes the hierarchical learning to correlate the connections between different modalities. Deep-SM [19] uses deep semantic matching to retrieve different modalities with multi-labels. The supervised DSCMR [7] learns the features by minimizing the discriminative loss between the label and common spaces. CMDN, Deep-SM and DSCMR are deep learning based methods.

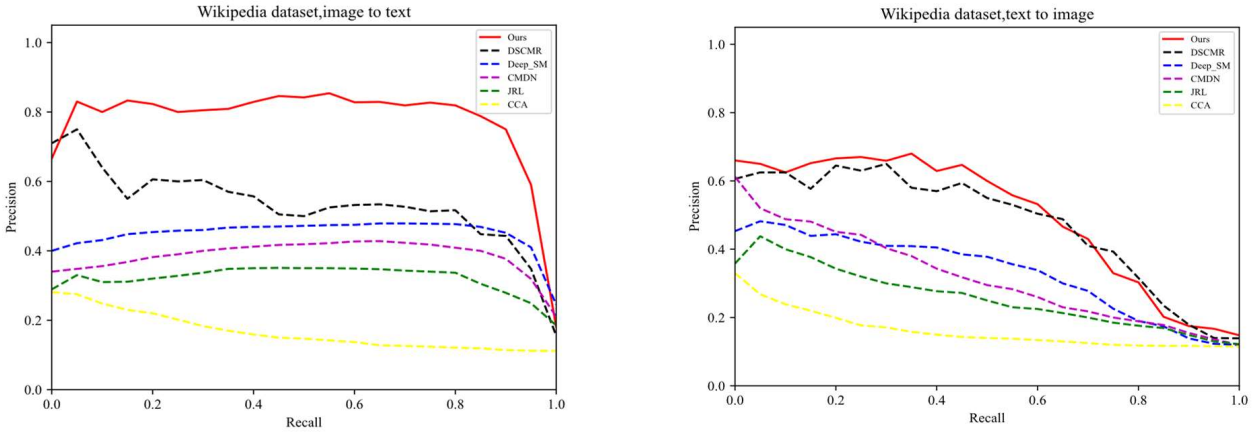
Tables 1 and 2 show the mAP values on the Wikipedia and Pascal Sentence datasets, respectively. Correspondingly, Figures 9 and 10 show the PR curves. The experimental results verify that our method outperforms the best-of-the-art methods.

Table 1. The mAP values of the cross-modal retrieval performance on the Wikipedia dataset.

Method	Task		
	Image to Text	Text to Image	Average
CCA	0.176	0.178	0.177
JRL	0.344	0.277	0.311
CMDN	0.393	0.325	0.359
Deep-SM	0.458	0.345	0.402
DSCMR	0.487	0.429	0.458
Ours	**0.494**	**0.437**	**0.466**

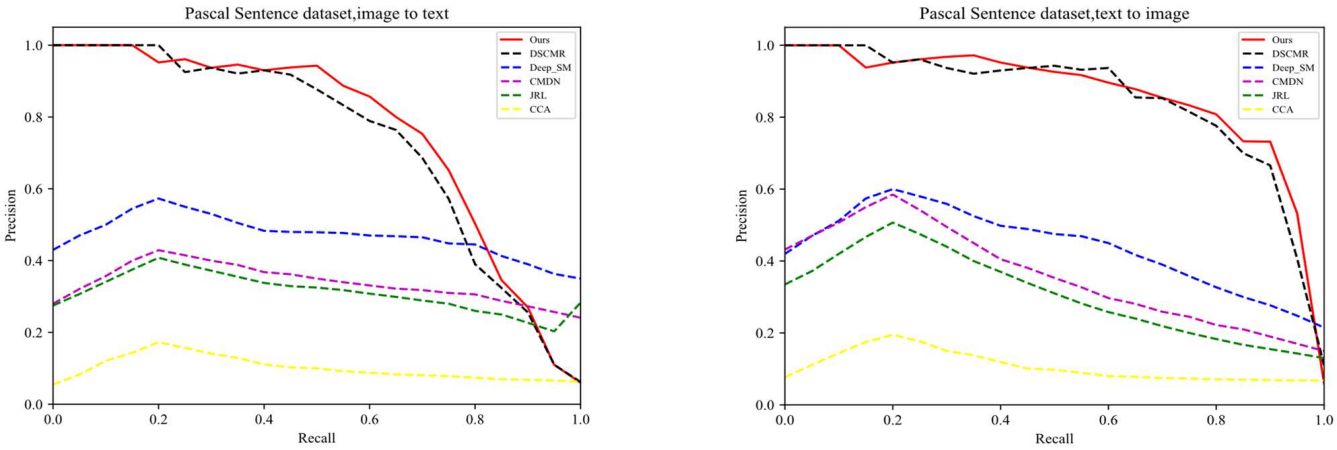
Table 2. The mAP values of the cross-modal retrieval performance on the Pascal Sentence dataset.

Method	Task		
	Image to Text	Text to Image	Average
CCA	0.110	0.116	0.113
JRL	0.300	0.286	0.293
CMDN	0.334	0.333	0.334
Deep-SM	0.440	0.414	0.427
DSCMR	0.688	0.704	0.696
Ours	**0.698**	**0.710**	**0.704**



(**a**) (**b**)

Figure 9. The PR curves of the cross-modal retrieval performance on the Wikipedia dataset. (**a**) The image to text task on the Wikipedia dataset. (**b**) The text to image task on the Wikipedia dataset.



(**a**) (**b**)

Figure 10. The PR curves of the cross-modal retrieval performance on the Pascal Sentence dataset. (**a**) The image to text task on the Pascal Sentence dataset. (**b**) The text to image task on the Pascal Sentence dataset.

CCA only uses the mutual relationship to study the correlation between two groups, so it cannot understand the class labels' high-level semantic information. As a result, CCA has a weak ability to discriminate the samples in the common space. JRL learns different modal sparse projection matrices and utilizes unlabeled data to improve the diversity of the training data. CCA and JRL use traditional methods to correlate different modal connections, which cannot make full use of samples' semantic information. To solve the above-mentioned problem, the deep learning mechanism is employed to further improve cross-modal retrieval performance. CMDN hierarchically combines the intra- and inter-modal features, and uses a two-level network strategy to learn the cross-modal correlations. However, the different modal distributions are diverse, and CMDN does not align the generated features' distribution. Deep-SM extracts visual features by a pre-trained network, and adopts a deep semantic matching method to achieve the cross-modal retrieval of the samples with multiple labels. However, Deep-SM does not take the same modal correlation into account. The supervised DSCMR learns the discriminative features and minimizes the discriminative loss between the label and common spaces. Moreover, DSCMR adopts a weight-sharing strategy to reduce the differences between different the modal high-level semantic information. However, DSCMR does not consider the intra-modal correlation during the training procedure. In this paper, the proposed method uses the adversarial network to jointly model the heterogeneous data and generate their feature representations in the common space, which can reduce the differences between the different modalities' feature distributions. Furthermore, it preserves both the inter- and intra-modal triplet similarity relationship. This measure can avoid retrieving negative samples, which may have similar intra-modal feature representations, and boost robustness to noise. In each modal space, a linear projection function is built to classify the generated feature, and the feature's prediction label is required to be consistent with the sample's label. As a result, the generated feature can correctly represent the sample's semantic information. Thanks to the above measures, the cross-modal retrieval task can be achieved directly based on different modal features' similarity relationship. Finally, the experimental results show that the proposed method is better than the state-of-the-art algorithms.

4.5. The Ablation Study on Constraints

In this paper, to guarantee the cross-modal retrieval performance, we designed three loss functions: (1) the cross-modal learning loss (L_{Ret}), which simultaneously preserved the inter- and intra-modal similarity relationship; (2) the discrimination loss (L_{Dis}), which aimed to preserve the semantic information and improve the intra-modal discriminative ability during generating features; and (3) the cross-modal adversarial loss (L_{Adv}), which aligned different modalities' feature distributions in the common space. To verify the effect of the above loss functions in the cross-modal retrieval task, we conducted the ablation study on the Wikipedia and Pascal Sentence datasets, and the comparison algorithms are shown in Table 3. The final experimental results are shown in Tables 4 and 5, and in Figures 11 and 12.

Table 3. Comparative algorithms in the ablation experiments.

The Ablation Algorithms	The Objective Functions
NoL_{Dis}	No Discrimination loss L_{Dis}
NoL_{Ret}	No Cross-modal learning loss L_{Ret}
NoL_{Adv}	No Cross-modal adversarial loss L_{Adv}
$OnlyL_{Dis}$	Only Discrimination loss L_{Dis}
$OnlyL_{Ret}$	Only Cross-modal learning loss L_{Ret}
$OnlyL_{Adv}$	Only Cross-modal adversarial loss L_{Adv}

Table 4. The mAP scores of the cross-modal retrieval performance on the Wikipedia dataset.

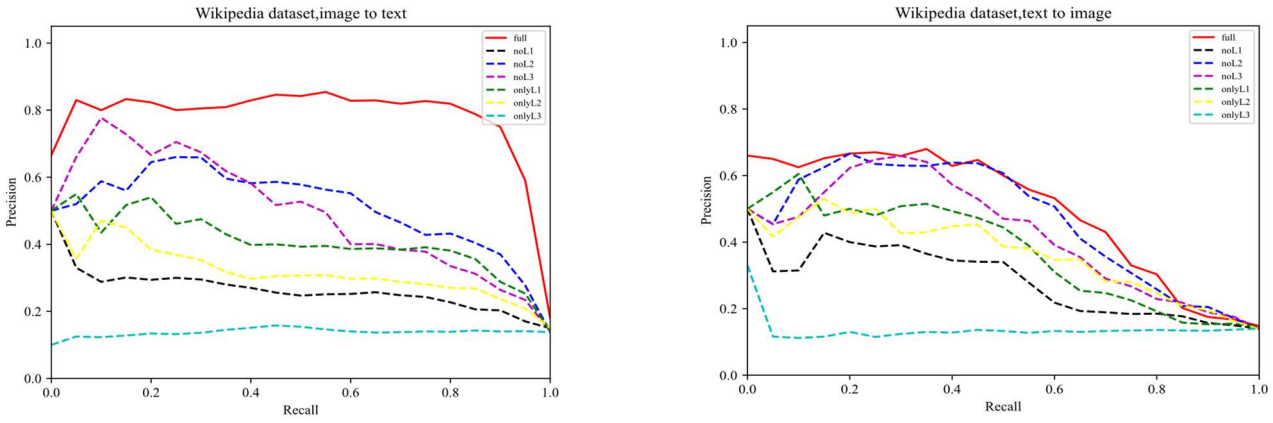
Method	Task		
	Image to Text	Text to Image	Average
NoL_{Dis}	0.279	0.270	0.274
NoL_{Ret}	0.441	0.405	0.423
NoL_{Adv}	0.441	0.400	0.421
$OnlyL_{Dis}$	0.439	0.405	0.422
$OnlyL_{Ret}$	0.357	0.337	0.347
$OnlyL_{Adv}$	0.120	0.123	0.122
Full	**0.494**	**0.437**	**0.466**

Table 5. The mAP scores of the cross-modal retrieval performance on the Pascal Sentence dataset.

Method	Task		
	Image to Text	Text to Image	Average
NoL_{Dis}	0.101	0.328	0.215
NoL_{Ret}	0.595	0.690	0.642
NoL_{Adv}	0.662	0.670	0.666
$OnlyL_{Dis}$	0.665	0.678	0.672
$OnlyL_{Ret}$	0.649	0.649	0.649
$OnlyL_{Adv}$	0.115	0.091	0.103
Full	**0.698**	**0.710**	**0.704**

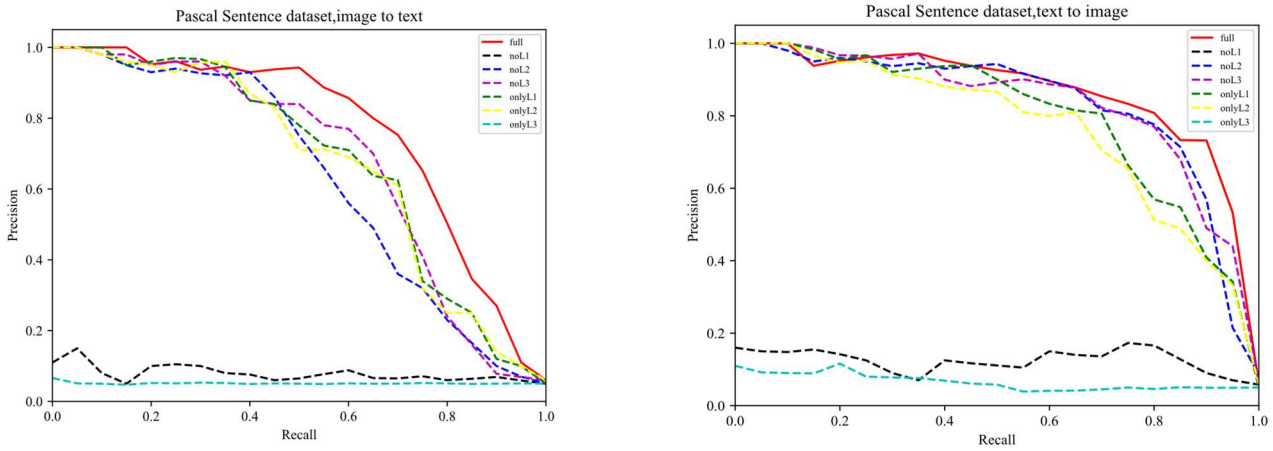
The experimental results show that the deep adversarial triplet similarity preserving cross-modal retrieval algorithm achieved the best performance on both datasets. This means all three loss functions played important roles in achieving the cross-modal retrieval task. When we generated the feature using the end to end network, L_{Dis} could preserve the semantic information. Moreover, L_{Dis} ensured the model could discriminate the samples belonging to different categories in the common space. Thus, the performance of $OnlyL_{Dis}$ was better than NoL_{Dis}. $OnlyL_{Adv}$ had a poor performance, because it only aligned different modal feature distributions, while ignoring preserving the similarity relationship among the intra-modal samples. In the common space, the cross-modal learning module minimized the distance between the same category samples and maximized the distance between

the different categories' samples. Assisted by the cross-modal learning module, we could return the samples similar to the query sample as the retrieval results.



(**a**) (**b**)

Figure 11. The PR curves of cross-modal retrieval performance on the Wikipedia dataset. (**a**) The image to text task on the Wikipedia dataset. (**b**) The text to image task on the Wikipedia dataset.



(**a**) (**b**)

Figure 12. The PR curves of cross-modal retrieval performance on the Pascal Sentence dataset. (**a**) The image to text task on the Pascal Sentence dataset. (**b**) The text to image task on the Pascal Sentence dataset.

4.6. The Ablation Study on Triplet Similarity Preserving Constraint

In this paper, to preserve the similarity relationship among samples, we designed the triplet similarity preserving objective function L_{Ret}, which included three parts, the inter-modal triplet similarity preserving constraint L_O, the image intra-modal triplet similarity preserving constraint L_v, and the text intra-modal triplet similarity preserving constraint L_t. During the training process, we learned different modal features by simultaneously minimizing the values of L_O, L_v, and L_t in the common space.

To further illustrate the importance of L_O, L_v, and L_t, we separately conducted the ablation experiments on the Pascal Sentence and Wikipedia datasets. Table 6 shows the final experimental results, which included the image to text and the text to image.

Table 6. The mAP values of the cross-modal retrieval results with minimal L_O or L_{Ret}.

Datasets	Task	Minimal L_O	Minimal L_{Ret}
Pascal Sentence	Image to Text	0.688	**0.698**
	Text to Image	0.701	**0.710**
	Average	0.694	**0.704**
Wikipedia	Image to Text	0.488	**0.494**
	Text to Image	0.417	**0.437**
	Average	0.453	**0.466**

The experimental results show that the cross-modal retrieval performance with minimal L_{Ret} was better than that with minimal L_O. By minimizing L_O, we could only guarantee the triplet similarity relationship among the different modal samples. Unfortunately, the intra-modal samples that belonged to the different categories may have similar features without preserving the intra-modal triplet similarity relationship. As a result, the cross-modal retrieval task may return the negative samples as the nearest neighbors. In contrast, by minimizing the value of L_{Ret}, we could simultaneously preserve the inter- and intra-modal triplet similarity relationship. This could guarantee that the distances among the inter- and intra-modal similar samples were smaller than those among the dissimilar samples. Thus, the inter- and intra-modal nearest neighbors were assigned similar features. Finally, we could improve the cross-modal retrieval performance assisted by L_{Ret}.

4.7. The Noise Robustness Experiments

In this section, we aimed to verify that the proposed triplet similarity relationship preserving constraint could boost the algorithm's robustness to noise. We randomly generate uniform noise, and separately put them into the Pascal Sentence and Wikipedia datasets. The cross-modal retrieval experiments are shown in Table 7 and Figure 13. We removed the triplet similarity relationship constraint (L_{Ret}) from the comparative algorithm. The final experimental results showed that the proposed triplet similarity preserving constraint L_{Ret} could effectively boost the algorithm's robustness to noise.

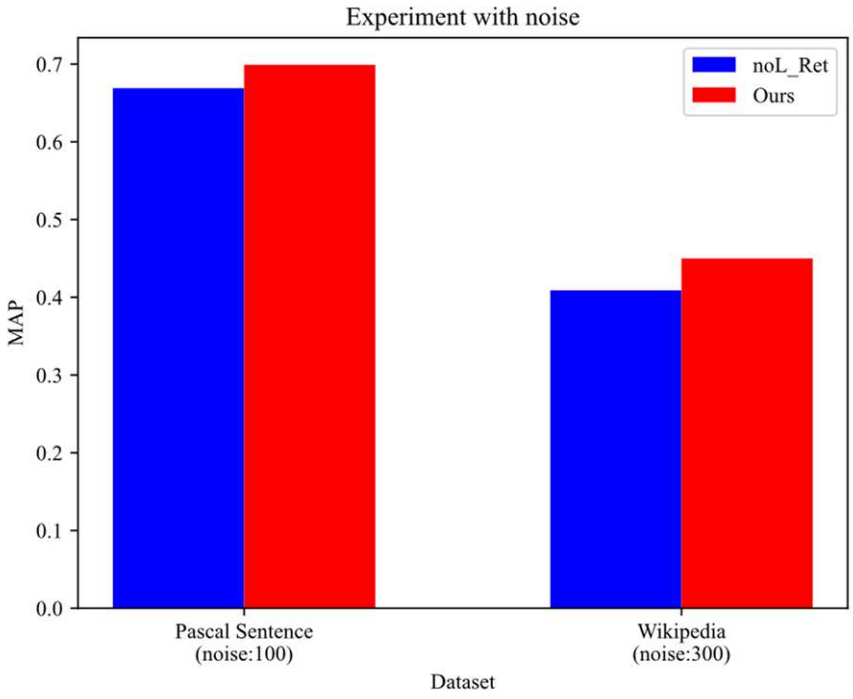


Figure 13. The mean mAP value of different cross-modal retrieval tasks on the Pascal Sentence and Wikipedia datasets, which contain 100 and 300 noise samples, respectively.

Table 7. The mAP values of the cross-modal retrieval performance on the Pascal Sentence and Wikipedia datasets, which contain 100 and 300 noise samples, respectively.

Datasets	Tasks	Without L_{Ret}	The Proposed Algorithm
Pascal Sentence (100 noise samples)	Image to Text	0.669	**0.690**
	Text to Image	0.668	**0.707**
	Average	0.669	**0.699**
Wikipedia (300 noise samples)	Image to Text	0.437	**0.481**
	Text to Image	0.380	**0.419**
	Average	0.409	**0.450**

5. Conclusions

The existing cross-modal algorithms do not align different modal feature distributions in the common space. Moreover, they do not take the similarity relationship preserving among the intra-modal samples into consideration. To solve these problems, we propose a novel cross-modal retrieval algorithm. We use the adversarial networks to generate different modal features and to align their distributions in the common space. To preserve the similarity relationship among the intra-modal samples, we establish the triplet similarity relationship preserving function to minimize the distance between the same category samples and to maximize the distance between the different categories' samples. This measure can avoid retrieving negative samples caused by noise interference and improves the robustness of the algorithm. During the training process, we utilize the linear function to project the generated features into different classes, and require the prediction label of the generated feature be as the same as the sample's label. Thus, it can minimize the semantic information loss while learning the features. To verify the cross-modal performance of the proposed method, we conduct comparative experiments on two widely used benchmark datasets—the Wikipedia and Pascal Sentence datasets. The final experimental results demonstrate that our proposed method achieves the best cross-modal retrieval performance.

Author Contributions: Conceptualization, Z.W. and G.L.; methodology, Z.W. and G.L.; software, X.Y. and G.L.; validation, S.X., X.Y. and C.F.; formal analysis, Z.W. and N.W.; investigation, C.F. and S.X. resources, F.S. and X.Y.; data curation, C.F.; writing—original draft preparation, G.L.; writing—review and editing, Z.W.; visualization, N.W. and F.S.; supervision, Z.W.; project administration, Z.W.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 61841602; the Natural Science Foundation of Shandong Province of China, grant number ZR2018PF005 and ZR2021MF017; the Youth Innovation Science and Technology Team Foundation of Shandong Higher School, grant number 2021KJ031; and the Fundamental Research Funds for the Central Universities, JLU, grant number 93K172021K12.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: The authors express their gratitude to the institutions that supported this research: Shandong University of Technology (SDUT) and Jilin University (JLU).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Z.; Lu, H.; Fu, H. Image-text bidirectional learning network based cross-modal retrieval. *Neurocomputing* **2022**, *483*, 148–159. [CrossRef]
2. Cai, L.; Zhu, L.; Zhang, H. DA-GAN: Dual Attention Generative Adversarial Network for Cross-Modal Retrieval. *Future Internet* **2022**, *14*, 43. [CrossRef]
3. Li, Z.; Xu, X.; Zhang, D.; Zhang, P. Cross-modal hashing retrieval based on deep residual network. *Comput. Syst. Sci. Eng.* **2021**, *36*, 383–405. [CrossRef]
4. Zhang, H.; Koh, J.Y.; Baldridge, J. Cross-Modal Contrastive Learning for Text-to-Image Generation. *arXiv* **2021**, arXiv:2101.04702.

5. Zhen, W.; Nannan, W.; Xiaohan, Y. Deep Learning Triplet Ordinal Relation Preserving Binary Code for Remote Sensing Image Retrieval Task. *Remote Sens.* **2021**, *13*, 4786. [CrossRef]
6. Zhen, W.; Fuzhen, S.; Longbo, Z. Minimal residual ordinal loss hashing with an adaptive optimization mechanism. *Eurasip. J. Image Video Proc.* **2020**, 10.
7. Zhen, L.; Hu, P.; Wang, X. Deep Supervised Cross-Modal Retrieval. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
8. Wang, H.; Sahoo, D.; Liu, C. Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st Internetional Conference on Internetional Conference on Machine Learning, Beijing, China, 21–26 June 2014; Volume 32.
11. Hotelling, H. Relations between two sets of variants. *Biometrika* **1935**, *28*, 312–377.
12. Li, D.; Dimitrova, N.; Li, M. Multimedia content processing through cross-modal association. In Proceedings of the Multimedia 03: Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, 2–8 November 2003.
13. Zhai, X.; Peng, Y.; Xiao, J. Learning Cross-Media Joint Representation With Sparse and Semisupervised Regularization. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 965–978. [CrossRef]
14. Peng, Y.; Qi, J. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–24. [CrossRef]
15. Ngiam, J.; Khosla, A.; Kim, M. Multimodal Deep Learning. In Proceedings of the International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009.
16. Wang, W.; Yang, X.; Ooi, B.C. Effective deep learning-based multi-modal retrieval. *VLDB J.* **2016**, *25*, 79–101. [CrossRef]
17. Andrew, G.; Arora, R.; Bilmes, J. Deep Canonical Correlation Analysis. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
18. Wang, W.; Arora, R.; Livescu, K. On deep multi-view representation learning. In Proceedings of the 32nd International Conference on International Conference Machine Learning, Lille, France, 6–11 July 2015; Volume 37.
19. Wei, Y.; Yao, Z.; Lu, C. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE Trans. Cybernetics* **2017**, *47*, 449–460. [CrossRef] [PubMed]
20. Jian, W.; He, Y.; Kang, C. Image-Text Cross-Modal Retrieval via Modality-Specific Feature Learning. In Proceedings of the 5th ACM on Internetional Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015.
21. Wang, B.; Yang, Y.; Xu, X. Adversarial Cross-Modal Retrieval. In Proceedings of the 25th ACM on International Conference on Multimedia, New York, NY, USA, 23–27 October 2017.
22. Zhai, D.; Chang, H.; Shan, S. Multiview Metric Learning with Global Consistency and Local Smoothness. *ACM Trans. Intell. Syst. Technol.* **2012**, *3*, 1–22. [CrossRef]
23. Sharma, A.; Kumar, A.; Daume, H. Generalized Multiview Analysis: A discriminative latent space. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
25. Wei, X.; Zhang, T.; Li, Y. Multi-Modality Cross Attention Network for Image and Sentence Matching. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
26. Zhou, R.; Shen, Y.D. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
27. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
28. Rashtchian, C.; Young, P.; Hodosh, M. Collecting image annotations using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010.
29. Pereira, J.C.; Coviello, E.; Doyle, G. On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 521–535. [CrossRef] [PubMed]
30. Peng, Y.; Xin, H.; Qi, J. Cross-media shared representation by hierarchical learning with multiple deep networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016.
31. Rasiwasia, N.; Pereira, J.C.; Coviello, E. A New Approach to Cross-Modal Multimedia Retrieval. In Proceedings of the 18th ACM International Conference on Multimedea, New York, NY, USA, 25–29 October 2010.