X-GACMN: An X-Shaped Generative Adversarial Cross-Modal Network with Hypersphere Embedding

Weikuo Guo[1], Jian Liang[2,3], Xiangwei Kong[4(✉)], Lingxiao Song[2], and Ran He[2,3]

[1] Dalian University of Technology, Dalian, China
guoweikuo@mail.dlut.edu.cn
[2] University of Chinese Academy of Science (UCAS), Beijing, China
{jian.liang,lingxiao.song,rhe}@nlpr.ia.ac.cn
[3] CRIPAC and NLPR, CASIA, Beijing, China
[4] Zhejiang University, Hangzhou, China
kongxiangwei@zju.edu.cn

Abstract. How to bridge heterogeneous gap between different modalities is one of the main challenges in cross-modal retrieval task. Most existing methods try to tackle this problem by projecting data from different modalities into a common space. In this paper, we introduce a novel X-Shaped Generative Adversarial Cross-Modal Network (X-GACMN) to learn a better common space between different modalities. Specifically, the proposed architecture combines the process of synthetic data generation and distribution adapting into a unified framework to make sure the heterogeneous modality distributions similar to each other in the learned common subspace. To promote the discriminative ability, a new loss function that combines intra-modality angular softmax loss and cross-modality pair-wise consistent loss is further imposed on the common space, hence the learned features can well preserve both inter-modality structure and intra-modality structure on a hypersphere manifold. Extensive experiments on three benchmark datasets show the effectiveness of the proposed approach.

Keywords: Cross-modal retrieval · Generative adversarial network · Hypersphere embedding

1 Introduction

With the help of well-annotated large scale datasets and advancing machine learning techniques, the majority computer vision and pattern recognition tasks have achieved impressive performance, such as machine translation [32], image classification [44], and object detection [13]. However, real-world information often presented in more complex ways at the same time. Tasks that are aiming at solving more complicated problems such as image captioning [18] and visual

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11365, pp. 513–529, 2019.
https://doi.org/10.1007/978-3-030-20873-8_33

question answering [11] always involve more than one modality data form. Over the past few years, multi-modal learning has become a super hot topic with the increase of massive multi-modal data [37]. To make computing equipment understand the world better, representing and matching data from different modalities is crucial and remains challenging.

Data from different modalities are always quite different from each other. Taking image modality and text modality as an example, images are constituted by pixels and show more details while texts are presented in the form of word sequences which contain high-level semantic information. Such differences between different modalities are collectively known as the heterogeneous gap which is one of the most challenging problems to solve in cross-modal retrieval task. Many of existing cross-modal retrieval works try to solve this problem by projecting different modality data into a common space in which similarity between different modality data can be quantitatively measured. Various cross-modal mapping methods [3,12,23,34] with all kinds of common space constraint losses emerge in recent years have achieved noticeable improvement on the cross-modal retrieval task. However, these methods still suffer from the lacking of effective constraint in the common space, which declines the cross-modal retrieval performance.

To make the learned features intra-modality discriminative, the softmax loss is widely used in cross-modal retrieval task. However, softmax loss only learns separable features that are not discriminative enough. Previous works [23,34,39] combine softmax loss with other Euclidean distance based constraints to enhance the discrimination power of features. However, the features learned with original softmax loss have natural angular distribution. Directly combining Euclidean distance based constraint with softmax loss may destroy such angular distribution. To take full advantage of the angular distribution of the original softmax loss, [16,17] make efforts to learn angular distributed features with A-softmax or L-softmax and achieve success in face recognition task.

Inspired by these works above, in this paper, we leverage angular constraint in cross-modal retrieval task to learn angular distributed common space representation. By constraining the hypersphere embedding metric for each modality to be the same and adding cross-modality pair-wise consistent to the original A-softmax, the similarity between different modality representations can be evaluated on a uniformed hypersphere manifold. The proposed cross-modal A-softmax abandon the time-consuming negative instance sampling process in triplet ranking based constraint, which makes it has high computational efficiency and not relying too much on the annotation of data. Besides, to the best of our knowledge, this work is the first work that tries to solve cross-modal retrieval problem with angular constraint, and the experimental result shows the effectiveness of the newly designed cross-modal A-softmax method.

However, the task of cross-modal retrieval expects the learned features to be not only intra-modality discriminative but also inter-modality coherent. Therefore, we should find another constraint to narrow the heterogeneous gap between different modalities. The emergence of Generative Adversarial Networks (GANs)

[7] brings new ideas to many tasks. With the development of GANs, computing equipment can generate different kinds of data adversely. Attempts like generating images with captions [28], and generating captions with images [18] have already proved to be possible. Although these works have different tasks from cross-modal retrieval, they show new possibilities of cross-modal matching. In addition, to generate synthetic data, adversarial training is also widely used in tasks like domain adaption [33] because it can make distributions of the learned features becoming close to each other. The task of cross-modal retrieval also requires that the learned features from different modalities have the same distribution, which makes people intuitively think to use GANs to solve the cross-modal problem. Actually, attempts have already been made in the previous works [23,34] and some of them achieve impressive performance.

Our proposed method tries to combine the process of synthetic data generation and distribution adapting into a unified adversarial training framework to make the learned common space representations to be more modality invariant. A novel X-shaped Generate Adversarial Network (X-GACMN) architecture as shown in Fig. 1 is designed to achieve this. In the proposed X-GACMN model, we first assume that there exists an intermediate state in the middle of the process of cross-modality synthetic data generation. By constraining this intermediate state with adversarial training, we can obtain a common space for different modalities and the heterogeneous gap can be implicitly narrowed.

The contribution of this paper is mainly threefold,

1. An X-shaped generative adversarial cross-modal network (X-GACMN) is designed for cross-modality matching to ensure inter-modality coherent. With two generators to form an information loop and three discriminators to constrain the feature distribution, the correlation between different modalities is maximized and the heterogeneous gap is narrowed.
2. A novel common space angular constraint is applied to learn more intra-modality discriminative features in the common space. With the cross-modal A-softmax constraint, the learned common representations are angularly distributed and can be measured on a hypersphere manifold effectively.
3. Experimental results on three public benchmarks show that the proposed X-GACMN achieves competitive results compared with other state-of-the-art cross-modal retrieval approaches.

2 Related Work

2.1 Cross-Modal Retrieval with GAN

Cross-modal retrieval is the most common and basic task among tasks involve more than one modality. Briefly speaking, cross-modal retrieval aims to bridge the heterogeneous gap between different modalities. Most of the existing cross-modal retrieval method try to solve this by learning common representation and can be divided into two broad categories according to the type of the target representation. Binary representation learning is also called hashing method [3,12],

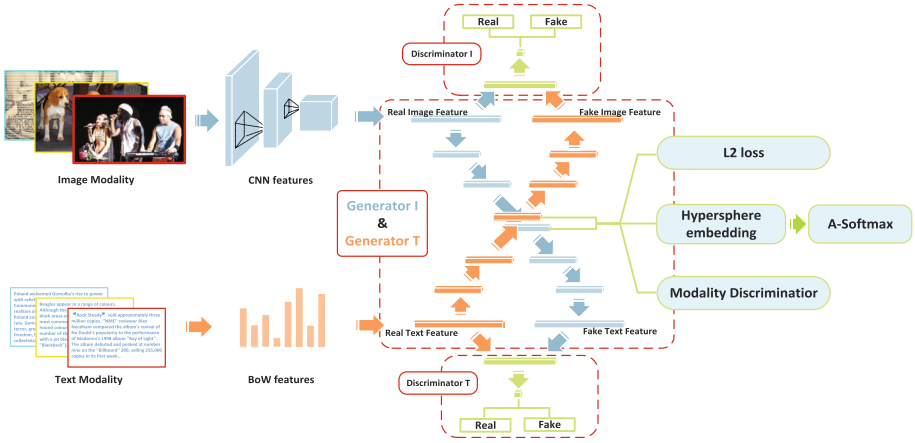

Fig. 1. The architecture of the proposed X-GACMN.

which aims to project original data or features from different modalities into a common hamming space. In the hamming space, the similarity between binary representations can be measured with hamming distance. These hashing methods have high computational efficiency, but sometimes at the cost of retrieval accuracy (effectiveness). The method this paper proposed falls into the other category called real-valued representation learning. Real value representing learning as [37] summarized can be divided into four subclasses: unsupervised [10,20,31], pairwise [25,40,41,43], ranking-based [5,8,19,30], and supervised [6,35,36,38,42] ones. Our approach is a supervised method with generative adversarial networks. Such combination can be seen in some recently published works [9,23,34]. In [34], a modality discriminator is directly applied to the learned multi-modal representations to discriminate generated representations from different modalities. The representations from different modalities get close to each other with adversarial learning. [23] involves intra-modality reconstruction constraint through appending discriminators to adversarially make reconstructed features getting close to original ones. Similar to our approach, the method [9] proposed also tries to use data from one modality to generate data from other modality. But their approach tries to use raw images to generate sentences whose performance could be constrained by the scale of the dataset.

Following but not limited to the works mentioned above, our X-GACMN has two generators to perform cross-modal generation. Three discriminators are appended to image feature space, text feature space and the learned common representation space to constrain cross-modal reconstruction and common representation learning process respectively.

2.2 Hypersphere Embedding

The idea of hypersphere embedding is first proposed in [16], and originally designed for face recognition. By modifying softmax loss to A-softmax loss, the original features can be constrained with an angular margin and obtain better recognition performance.

The original softmax loss is designed for classification tasks. Many works try to improve the original softmax loss to make it learn more discriminative features. L-softmax loss [17] also involves the concept of angle but it doesn't normalize the weights, thus the learned features are not constrained to hypersphere manifold and not suitable for open-set problems. Another modified method proposed in [39] combines softmax loss with Euclidean distance constraint by minimizing the distance between intra-class samples and the class center.

In cross-modal retrieval task, softmax loss is also widely used. Previous work like [23,34] combine softmax loss with Euclidean distance based triplet ranking loss to make the learned representations more semantically discriminative. However, it has been proved in [16,17] that the representations learned with original softmax loss have nature angular distribution, immediately combining Euclidean distance based constraint with softmax loss may destroy such angular distribution and cause bad influence.

In the proposed X-GACMN, we follow the above thinking and apply the angular constraint to cross-modal task and designed a cross-modal A-softmax to take full advantage of the angular constraint of softmax loss so as to enhance the retrieval performance.

3 Our Approach

As shown in Fig. 1, the network structure of the proposed X-GACMN is made up of two modules. The feature extracting module extracts original features $\mathcal{V}$ and $\mathcal{T}$ from image modality and text modality respectively. In this paper currently existing feature extracting methods such as CNN modal and BoW are applied. As for the feature projecting module, we will explain it in detail in the following subsections.

3.1 Notation

Without losing generality, we aim to conduct cross-modal representation learning on two modalities e.g., image and text. We assume there exists a multimodal training dataset which is composed of image-text pairs, denoted as $D = \{(v,t)_i\}_{i=1}^{n}$ where $(v,t)_i$ represents the i-th instance of image-text pair and n is the total number of instances in the dataset. In addition, each instance is assigned a semantic category label $\{c_i\}_{i=1}^{n}$. After feature learning phase, instances can be represented by original features, and the dataset converts to $D = \{\mathcal{V}, \mathcal{T}\}$ where $\mathcal{V} = \{v_{o1}, \ldots, v_{on}\} \in \mathbb{R}^{d_v \times n}$ and $\mathcal{T} = \{t_{o1}, \ldots, t_{on}\} \in \mathbb{R}^{d_t \times n}$ denote the original feature matrixes of image and text respectively and d_v, d_t are the dimensions of the original features.

The original features $v_o \in \mathcal{V}$ and $t_o \in \mathcal{T}$ may follow different complex distributions and have different kinds of statistical properties and dimensions. Thus it is hard to compare them directly. Our primary goal is to find a common subspace $\mathcal{S}$ in which the features after projection is comparable so that the similarity between different modalities can be calculated. The mapping functions for each modality can be formulated as $f_{\mathcal{V}}(v_o; \theta_{\mathcal{V}})$ and $f_{\mathcal{T}}(t_o; \theta_{\mathcal{T}})$. After projection, the original features are transformed into $\mathcal{S}_{\mathcal{V}} \in \mathbb{R}^{d_c \times n}$ and $\mathcal{S}_{\mathcal{T}} \in \mathbb{R}^{d_c \times n}$, where d_c represents the dimension of the common representation. The ultimate goal of the proposed approach X-GACMN is to make $\mathcal{S}_{\mathcal{V}}$ and $\mathcal{S}_{\mathcal{T}}$ modality-invariant and semantically discriminative. To achieve this goal, we apply X-shaped GAN structure to establish an information loop between different modalities and cross-modal A-softmax to maintain the underlying semantic information on a hypersphere manifold.

3.2 X-Shaped Generative Adversarial Cross-Modal Network

X-shaped architectures can be found in some recently published cross-modal researches [2,9]. These X-shaped architectures maximize the correlation between two modality-specific feature spaces by projecting data from one modality into the common representation space and then using the projected common representations to reconstruct data from the other modality. By minimizing the reconstruction loss, the correlation between the two pathways can be maximized. Just like these works, the proposed X-GACMN applies two cross-modal generators G_I and G_T to accomplish the task of cross-modal feature generation. The ultimate goal of these two generators is twofold: (1) to maximize the similarity between the generalized synthetic data and the real data. (2) to make the distribution of the learned common representations in the middle of each generator as close as possible. To accomplish this, three discriminators D_I, D_T and D_C are applied to image feature space, text feature space and common representation space respectively. By training the discriminators and generators iteratively, these two kinds of modal can beat each other with a minimax game, and finally, make the features in the aforementioned three kinds of feature space modality-invariant.

Each generator in the proposed X-GACMN modal is composed by an encoder and a decoder. Original features v_o and t_o are projected to common representations v_c and t_c with encoder $G_{I_{enc}}$ and $G_{T_{enc}}$ respectively. Then the reconstruction representations v_r and t_r can be captured with $G_{I_{dec}}$ and $G_{T_{dec}}$ respectively. The three discriminators are of two kinds. Discriminators D_I and D_T are synthetic data discriminators which are designed to discriminate generated synthetic data from the real ones. The adversarial loss of them can formally be defined as:

$$\mathcal{L}_{D_I} = -\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\log D_I(v_{o_i}; \theta_{D_I}) + \log\left(1 - D_I(v_{r_i}; \theta_{D_I})\right) \right) \tag{1}$$

$$\mathcal{L}_{D_T} = -\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\log D_T(t_{oi}; \theta_{D_T}) + \log\left(1 - D_T(t_{ri}; \theta_{D_T})\right) \right) \tag{2}$$

where n_{tr} denotes the total number of instances in training set, θ_{D_I} and θ_{D_T} are the parameters of D_I and D_T respectively.

Discriminators D_C is a modality discriminator which is designed to discriminate projected representations v_c and t_c in the common space. The adversarial loss of D_C can be defined as:

$$\mathcal{L}_{D_C} = -\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\log D_C \left(v_{ci}; \theta_{D_C} \right) + \log \left(1 - D_C \left(t_{ci}; \theta_{D_C} \right) \right) \right) \tag{3}$$

where θ_{D_C} are the parameters of D_C After discriminators D_I, D_T and D_C been optimized, generators G_I and G_T are optimized with θ_{D_I}, θ_{D_T} and θ_{D_C} fixed and the loss of them can be defined as following:

$$\mathcal{L}_{G_i} = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\log D_T \left(G_I \left(v_o; \theta_{G_I} \right) \right) + \log D_C \left(G_{I_{enc}} \left(v_o; \theta_{G_I} \right) \right) \right) \tag{4}$$

$$\mathcal{L}_{G_t} = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(\log D_I \left(G_T \left(t_o; \theta_{G_T} \right) \right) + \log D_C \left(G_{T_{enc}} \left(t_o; \theta_{G_T} \right) \right) \right) \tag{5}$$

where θ_{G_i} and θ_{G_t} are the parameters of G_i and G_t respectively.

3.3 Common Space Constraint

In this work, we introduce an angular constraint to cross-modal retrieval task and make efforts to adjust it to the multi-modal scenario. The angular constraint ensures the projected representations with different semantic labels be discriminative on a hypersphere manifold. By projecting representations from different modalities onto the same hypersphere manifold, the heterogeneous gap between different modalities can be further narrowed. The proposed angular constraint abandons the time consuming negative sampling and distance calculation process in the Euclidean distance based triplet constraint, which makes it not only preserve the angular feature distribution but also have high computational efficiency without relying too much on the annotation of data. More detailed descriptions are followed in the remainder of this section.

The widely used original softmax loss can be written as

$$\mathcal{L}_{softmax} = \frac{1}{N} \sum_i -\log \left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \right) \tag{6}$$

where N is the number of training instances and f is the posterior probabilities of input feature x_i. The hypothesis function can be represented as $f_j = W_j^T x_i + b_j$ and $f_{y_i} = W_{y_i}^T x_i + b_j$ where W_j^T and $W_{y_i}^T$ denotes the j-th and y_i-th column of the weight metric of the last fully connected layer in the CNN modal. We can rewrite Eq. (6) as follows in angular form:

$$\mathcal{L}_{softmax} = \frac{1}{N} \sum_i -\log \left(\frac{e^{\left\| W_{y_i}^T \right\| \left\| x_i \right\| \cos\left(\left\langle W_{y_i}^T, x_i \right\rangle \right) + b_{y_i}}}{\sum_j e^{\left\| W_j^T \right\| \left\| x_i \right\| \cos\left(\left\langle W_j^T, x_i \right\rangle \right) + b_j}} \right) \tag{7}$$

where $\langle W_*^T, x_i \rangle$ is the angle between feature x_i and W_*^T.

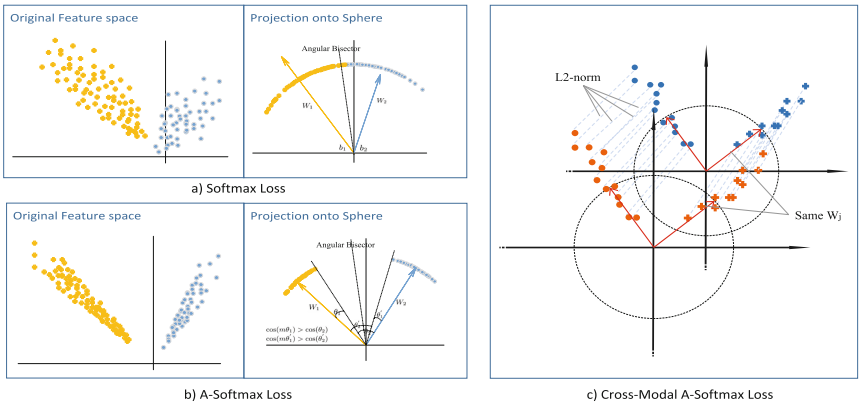

Fig. 2. A Comparison between softmax loss and A-softmax loss.

A-softmax makes ameliorate on the basis of the original softmax by firstly normalize $\left\|W_j^T\right\| = 1, \forall j$ and zero the biases, so that the original decision boundary for class i and class j can be presented in an angular margin form $\|x_i\| (\cos(\theta_i) - \cos(\theta_j))$, where θ_* is the angle between W_*^T and x_i. Secondly, A-softmax introduce a lower bound parameter m to quantitatively control the decision boundary and enhance the discrimination power. The decision boundary can be denoted as $\|x_i\| (\cos(m\theta_i) - \cos(\theta_j))$ and $\|x_i\| (\cos(\theta_i) - \cos(m\theta_j))$ for each class respectively. These two steps makes the features learned with A-softmax have angular margin and semantically discriminative. The A-softmax loss can be formulated as:

$$\mathcal{L}_{A-softmax} = \frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \psi\left(\left\langle W_{y_i}^T, x_i \right\rangle\right)}}{e^{\|x_i\| \psi\left(\left\langle W_{y_i}^T, x_i \right\rangle\right)} + \sum_{j \neq y_i} e^{\|x_i\| \cos\left(\left\langle W_j^T, x_i \right\rangle\right)}} \right) \tag{8}$$

where

$$\begin{aligned} &\psi\left(\left\langle W_{y_i}^T, x_i \right\rangle\right) = (-1)^k \cos(m \langle W_{y_i}, x_i \rangle) - 2k, \\ &\langle W_{y_i}, x_i \rangle \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m} \right], k \in [0, m-1] \end{aligned} \tag{9}$$

$\psi\left(\left\langle W_{y_i}^T, x_i \right\rangle\right)$ is a monotonically decreasing angle function which is generalized by expanding the definition range of $\cos\left(\left\langle W_{y_i}^T, x_i \right\rangle\right), \left\langle W_{y_i}^T, x_i \right\rangle \in \left[0, \frac{\pi}{m}\right]$. $m \geqslant 1$ is an integer parameter that controls the size of angular margin. With bigger m narrower angular margin can be obtained. The difference between the original softmax loss and A-softmax can be seen in Fig. 2a and b.

The A-softmax has an intuitive hypersphere interpretation. Because A-softmax loss requires $\|W_j\| = 1, b_j = 0$, the original features are projected to a

hypersphere manifold, on which the similarity between instances can be quantitatively evaluated by angle or the length of hyperarc. In our cross-modal task, we not only need to make sure instances with different semantic labels be discriminative but also expect items from different modalities can be quantitatively evaluated on the same manifold. Hance for different modalities, we use the same W_j to make sure the original features are projected to the same hypersphere manifold so that the heterogeneous gap between different modalities can be further narrowed. To maximize the correlation between two modalities while at the same time not destroying the inner modality angular distribution, an additional l_2 norm loss is added as a pair-wise consistent constraint to the common space to ensure the representations belong to the same image-text pairs as close as possible. A sketch map of the proposed constraint can be seen in Fig. 2c.

3.4 Loss Function and Optimization

The final objective functions for the generators G_I and G_T can be written as:

$$\mathcal{L}_{G_I} = \lambda_1 \mathcal{L}_{l_2} + \lambda_2 \mathcal{L}_{G_i} + \lambda_3 \mathcal{L}_{A-Softmax_i}, \tag{10}$$

$$\mathcal{L}_{G_T} = \lambda_1 \mathcal{L}_{l_2} + \lambda_2 \mathcal{L}_{G_t} + \lambda_3 \mathcal{L}_{A-Softmax_t}, \tag{11}$$

where λ_1, λ_2 and λ_3 are the weights of l_2 loss, generation loss and A-softmax loss respectively.

The process of optimizing the feature representation is conducted by optimizing the generator loss and the discriminator loss iteratively. The optimization goal of these two stage are opposite, which makes it a minimax game [7] of two sub-processes:

$$\begin{aligned} &\arg\min \left(\mathcal{L}_{G_I} \left(\hat{\theta}_{G_I}, \hat{\theta}_{A-Softmax} \right) - \mathcal{L}_{D_T} - \mathcal{L}_{D_C} \right) \\ &\arg\min \left(\mathcal{L}_{G_T} \left(\hat{\theta}_{G_T}, \hat{\theta}_{A-Softmax} \right) - \mathcal{L}_{D_I} - \mathcal{L}_{D_C} \right) \end{aligned} \tag{12}$$

$$\begin{aligned} &\arg\max \left(\mathcal{L}_{G_I} - \mathcal{L}_{D_T} \left(\hat{\theta_{D_T}} \right) - \mathcal{L}_{D_C} \left(\hat{\theta_{D_C}} \right) \right) \\ &\arg\max \left(\mathcal{L}_{G_T} - \mathcal{L}_{D_I} \left(\hat{\theta_{D_I}} \right) - \mathcal{L}_{D_C} \left(\hat{\theta_{D_C}} \right) \right) \end{aligned} \tag{13}$$

The overall training procedure is presented in Algorithm 1.

4 Experiments

We conduct experiments on three widely-used cross-modal datasets including Wikipedia dataset [24], NUS-WIDE-10k dataset [1] and Pascal Sentence dataset [26]. Comparisons with other state-of-the-art methods on these three datasets verify the effectiveness of our proposed X-GACMN. Additional ablation study and visualization results are presented in the later part in order to dissect our method in detail.

Algorithm 1. Pseudo code of optimizing our X-GACMN.

Require: $\mathcal{V}_{tr}, \mathcal{T}_{tr}$: Training data from both modality; N: Batch size; m, λ_*: hyperparameters;

update until X-GACMN converges:

1: $v_{c_i}, v_{r_i}, t_{c_i}, t_{r_i}$ are generated by G_I and G_T respectively.
2: Calculate loss of D_I, D_T and D_C with equation (1), (2) and (3) respectively.
3: Optimize D_I, D_T and D_C with equation(13))
4: **for** K steps **do**
5: Calculate loss of G_I, G_T with equation (4) and (5)
6: Optimize G_I and G_T with equation(12)
7: **end for**
8: **return** Modal parameter $\theta_{G_I}, \theta_{G_T}$ and common space feature $f_{\mathcal{V}}\left(v_o; \theta_{\theta_{G_I}}\right)$ and $f_{\mathcal{T}}\left(t_o; \theta_{G_T}\right)$.

4.1 Datasets and Experimental Setup

Datasets. In this subsection, we briefly introduce the three datasets and the corresponding features in the experiment.

Wikipedia is a widely used dataset for cross-modal retrieval which consists of 2173 training image-text pair and 693 testing image-text pair annotated by 10 semantic labels. In some works, another dataset partition is used by separate the dataset into 1300 training pairs and 1566 testing pairs. In our experiment, we conduct experiments on both partition protocols. 4096-D VGG-19 [29] features and 1000-D BoW features are used for image modality text modality respectively. Besides, for a fair comparison with earlier methods, we also conduct experiments with 128-D SIFT features and 10-D LDA features for each modality.

Pascal Sentence contains 1000 images with 20 semantic labels. Each image is described by 5 sentences. We divide this dataset into 900 training instances and 100 testing instances as [22,34] did, and use the same 4096-D VGG-19 features and 1000-D BoW features for image and text modality respectively.

NUS-WIDE-10k is constructed by sampling 10,000 image text pairs from 10 largest categories of NUS-WIDE without overlaps. Following [22,34], 8000 training pairs and 1000 testing pairs are used in our experiment and 4096-D VGG-19 features and 1000-D BoW features are used for each modality.

Implementation Details. The proposed X-GACMN modal realize feature projection and reconstruction with G_I and G_T which are composed of 6 fully connected layers with tanh as active function. The numbers of hidden units in each network are $V \rightarrow 512 \rightarrow 100 \rightarrow 100 \rightarrow 100 \rightarrow 512 \rightarrow T$ for G_I and $T \rightarrow 512 \rightarrow 100 \rightarrow 100 \rightarrow 100 \rightarrow 512 \rightarrow V$ for G_T. In the middle of each generator, the 100 dimensional output is the common subspace representation to be learned.

Synthetic data discriminator D_I and D_T with structure $V(T) \rightarrow 2000 \rightarrow 2$ are appended to discriminate generated synthetic data from real ones. Modality discriminator D_C with structure $100 \rightarrow 50 \rightarrow 2$ is appended to the common feature space to discriminate learned common space features' modal. An angu-

lar softmax layer with the same parameters is appended to both modalities to constrain the learned features to obey angular distribution.

As for hyper-parameters of the modal. The batch size is set to 64 and m is set to 5. λ_1, λ_2 and λ_3 are set to 10, 5 and 100 respectively to make sure the scale of each item balance.

4.2 Experimental Results

Comparison with State-of-the-Art Methods. We first compare our X-GACMN with 13 state-of-the-art methods on three datasets. We choose CCA [10], CCA-3V [6], LCFS [36], JRL [42], JFSSL [35], PACMR [14], SM [27] and SPGCM [15] as traditional cross-modal retrieval methods, Multimodal-DBN [31], Bimodal-AE [20], Corr-AE [4], and CMDN [21], as deep learning based methods and ACMR [34] as GAN based method.

Table 1. Cross-modal retrieval comparison in terms of the mAP on Wikipedia dataset

Protocol	Methods	Shallow feature			Deep feature		
		i2t	t2i	Avg.	i2t	t2i	Avg.
1300/1566 [36]	CCA	0.255	0.185	0.220	0.267	0.222	0.245
	M-DBN	0.149	0.150	0.150	0.204	0.183	0.194
	Bimodal-AE	0.236	0.208	0.222	0.314	0.290	0.302
	SPGCM	0.265	0.207	0.236	0.390	0.362	0.376
	SM	0.260	0.242	0.251	0.475	0.389	0.432
	CCA-3V	0.275	0.224	0.249	0.437	0.383	0.410
	LCFS	0.279	0.214	0.246	0.455	0.398	0.427
	Corr-AE	0.280	0.242	0.261	0.402	0.395	0.398
	JRL	0.344	0.277	0.311	0.453	0.400	0.426
	PACMR	0.318	0.224	0.271	0.468	0.429	0.449
	JFSSL	0.306	0.228	0.267	0.428	0.396	0.412
	CMDN	-	-	-	0.488	0.427	0.458
	ACMR	0.316	0.227	0.272	0.477	0.435	0.456
	X-GACMN	**0.348**	**0.282**	**0.315**	**0.490**	**0.456**	**0.473**
2173/693 [27]	SPGCM	0.254	0.203	0.228	0.351	0.327	0.339
	SM	0.255	0.226	0.205	0.479	0.384	0.431
	LCFS	0.266	0.209	0.238	0.455	0.417	0.436
	PACMR	0.309	0.220	0.264	0.478	0.433	0.456
	ACMR	0.310	0.223	0.267	0.476	0.431	0.454
	X-GACMN	**0.326**	**0.241**	**0.284**	**0.501**	**0.435**	**0.468**

Tables 1 and 2 shows the experimental results in terms of mAP on Wikipedia dataset, Pascal Sentence dataset and the NUS-WIDE-10k dataset. From the

Table 2. Cross-modal retrieval comparison in terms of the mAP on Pascal Sentence dataset and NUS-WIDE-10k dataset

Methods	Pascal sentence			NUS-WIDE-10k		
	i2t	t2i	Avg.	i2t	t2i	Avg.
CCA	0.363	0.219	0.291	0.189	0.188	0.189
M-DBN	0.477	0.424	0.451	0.201	0.259	0.230
Bimodal-AE	0.456	0.470	0.458	0.327	0.369	0.348
LCFS	0.442	0.357	0.400	0.383	0.346	0.365
Corr-AE	0.489	0.444	0.467	0.366	0.417	0.392
JRL	0.504	0.489	0.496	0.426	0.376	0.401
CMDN	0.534	0.534	0.534	0.492	0.515	0.504
ACMR	**0.535**	0.543	0.539	0.447	0.505	0.476
X-GACMN	0.532	**0.547**	**0.540**	**0.501**	**0.526**	**0.514**

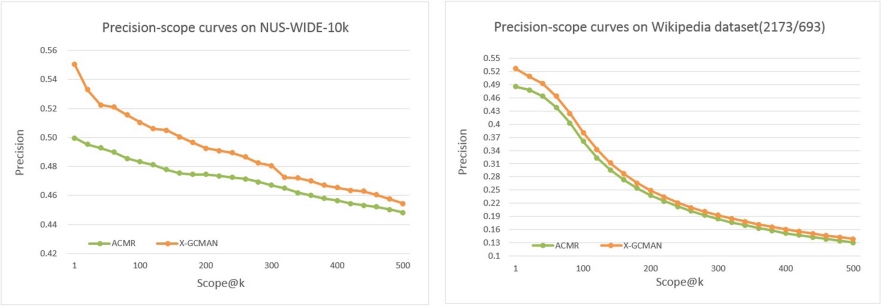

Fig. 3. Precision-scope curves on Wikipedia dataset and NUS-WIDE-10k dataset

results we have the following observations: (1) The proposed X-GACMN outperforms other methods with a big margin on these three datasets. Experimental results on these three datasets which have very distinct properties can testify the effectiveness of the proposed X-GACMN. (2) On Wikipedia dataset, we conduct experiments with shallow features and deep features on two different partition protocols. The performance of the proposed X-GACMN achieves best results with all these four different settings, which can prove the applicability of our method. (3) Compared with our best competitor ACMR which is also a GAN based method, our method obtained better results. This can preliminary shows the superiority of the proposed X-shaped GAN architecture and cross-modal A-softmax. More detailed comparison and analysis of these two methods are in the following subsections. (4) The performance on the Pascal Sentence dataset only be slightly improved, this is mainly because that the training data in Pascal Sentence dataset is limited in number and the X-GACMN model sufferers from

Table 3. Cross-modal retrieval comparison with different loss setting on Wikipedia dataset (1300/1766) in terms of the mAP@50

Methods	t2i	i2t	Avg.
X-GACMN without D_C	0.615	0.477	0.546
X-GACMN without D_I and D_T	0.618	0.482	0.550
X-GACMN without $\mathcal{L}_{A-Softmax}$	0.218	0.188	0.203
X-GACMN without $\mathcal{L}_{l_2}$	0.632	0.472	0.552
X-GACMN with $\mathcal{L}_{Softmax}$	0.598	0.467	0.533
Whole X-GACMN	**0.640**	**0.483**	**0.562**

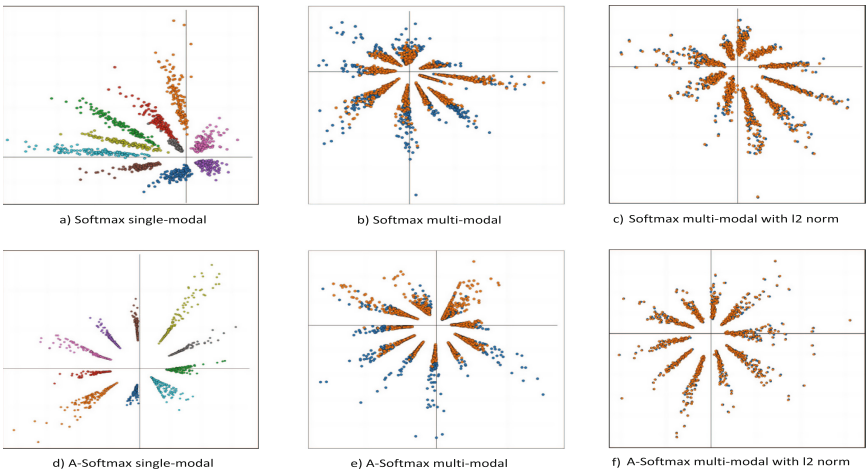

Fig. 4. The visualization result on Wikipedia dataset (1300/1766).

the overfitting problem, while in the large-scale dataset NUS-WIDE-10k, such problem is mitigated and the performance of our modal is sorted.

To further compare our method with the other GANs based method ACMR [34], we draw precision-scope curves on Wikipedia dataset and NUS-WIDE-10k dataset for additional comparison. The results can be seen in Fig. 3 From the curves we can see that our method outperforms ACMR with all scopes especially when the scopes are small. In real life retrieval scenario, we are more concern about the previous recalls, which means our method is significant in practical applications.

Ablation Study. In this section, we will discuss the effectiveness of each element in the X-GACMN modal. Table 3 summarizes the mAP@50 scores on Wikipedia dataset (2173/693) with different settings. To verify the effectiveness of the three adversarial training processes, we remove two kinds of discriminators respectively. From the first two lines, we can observe that the mAP@50 score drops when the synthetic data discriminators or the modality discriminator is missing,

which proves by training the modal adversely with the X-shaped architecture, a better common feature space can be learned.

Line 3, 4, 5 and 6 show the results with different common space constraint loss function. From line 3, we can see that without A-softmax loss the performance drops significantly, which is because the modal is trained unsupervisely without any semantic information. Line 4 shows the experimental results without l_2 norm loss, from the results we can see that the l_2 norm loss as a cross-modal pair-wise consistent constraint can improve the performance slightly. To compare the effectiveness of cross-modal A-softmax loss with the original softmax loss, we trained our model with original softmax and the results can be seen in line 5. From the results, we can see that cross-modal A-softmax is beneficial for the X-GACMN to learn more discriminative features.

Visualization of the Learned Feature Distribution. The A-softmax used in our X-GACMN has an intuitive hypersphere interpretation. To intuitively show the properties of the proposed common space constraint, we remove the last tanh active function and set the output dimension to 2, so that the learned features can be visualized in the two-dimensional space. The 2-D visualization of training feature distribution of Wikipedia dataset (2173/693) with different common space constraint can be seen in Fig. 4.

The first column of Fig. 4 shows the intra-modality distribution of common representations learned with softmax loss and A-softmax loss. From the visualization results, we can see that the features learned with the original softmax loss have natural angular distribution but sometimes not clear enough. Besides, the original softmax is designed for classification tasks which aim to find the best decision boundaries. Such a goal makes the cosine distance between features from different classes not necessarily smaller than features from the same class, which makes it not suitable for retrieval tasks. As for features learned with A-softmax, clear angular distributed margins between different classes can be seen. Such property ensures the learned features intra-modality discriminative.

The second and the third column of Fig. 4 shows the inter-modality visualization result. The difference between column 2 and column 3 is whether l_2 loss is applied or not. We can see that features obtained with A-softmax retain the inter-modality angular margin. This is because in our X-GACMN, features from different modalities are projected to a common hypersphere manifold with the same angular constraint. Another observation is by combining l_2 norm with A-softmax, the discrepancy between two different modalities is further diminished, which is because that by constraining distance of features belong to the same image-text pairs, the inter-modality structure can be preserved.

5 Conclusion

In this paper, we proposed a new X-shaped Generative Adversarial Cross-Modal Network (X-GACMN) to learn better common space representations for cross-modal retrieval. Firstly, the proposed X-GACMN designed an X-shaped GAN architecture to combine cross-modal synthetic data generation and distribution

adaption together with adversarial training. Secondly, the proposed X-GACMN for the first time exploited the angular constraint in cross-modal retrieval task to increase the discriminative ability of the learned features. With the X-shaped architecture and A-softmax, original features from different modalities are projected to a common hypersphere manifold on which the similarities between instances can be quantitatively evaluated by the magnitude of angle. Extensive experiments on three widely used cross-modal datasets and a detailed analysis of the experimental results demonstrate the effectiveness of our method.

Acknowledgement. We would like to thank anonymous reviewers for their helpful comments on the paper. This research was supported by the National Natural Science Foundation of China (NSFC) under Grant 61772111.

References

1. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national University of Singapore. In: Proceedings of the CIVR, pp. 48:1–48:9 (2009)
2. Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: Proceedings of the CVPR, pp. 4601–4611 (2017)
3. Erin Liong, V., Lu, J., Tan, Y.P., Zhou, J.: Cross-modal deep variational hashing. In: Proceedings of the ICCV, pp. 4077–4085 (2017)
4. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the ACM MM, pp. 7–16 (2014)
5. Frome, A., et al.: Devise: a deep visual-semantic embedding model. In: Proceedings of the NIPS, pp. 2121–2129 (2013)
6. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV **106**(2), 210–233 (2014)
7. Goodfellow, I.J., et al.: Generative adversarial nets. In: Proceedings of the NIPS, pp. 2672–2680 (2014)
8. Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. IEEE TPAMI **30**(8), 1371–1384 (2008)
9. Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: improving textual-visual cross-modal retrieval with generative models. In: Proceedings of the CVPR, pp. 7181–7189 (2018)
10. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
11. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: end-to-end module networks for visual question answering. In: Proceedings of the ICCV, pp. 804–813 (2017)
12. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of the CVPR, pp. 3270–3278 (2017)
13. Li, Y., Zhang, J., Huang, K., Zhang, J.: Mixed supervised object detection with robust objectness transfer. IEEE TPAMI **99**, 1–18 (2018)
14. Liang, J., Cao, D., He, R., Sun, Z., Tan, T.: Principal affinity based cross-modal retrieval. In: Proceedings of the ACPR, pp. 126–130 (2015)
15. Liang, J., He, R., Sun, Z., Tan, T.: Group-invariant cross-modal subspace learning. In: Proceedings of the IJCAI, pp. 1739–1745 (2016)

16. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep hypersphere embedding for face recognition. In: Proceedings of the CVPR, pp. 212–220 (2017)
17. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: Proceedings of the ICML, pp. 507–516 (2016)
18. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the CVPR, pp. 3242–3250 (2017)
19. Lu, X., Wu, F., Tang, S., Zhang, Z., He, X., Zhuang, Y.: A low rank structural large margin method for cross-modal ranking. In: Proceedings of the SIGIR, pp. 433–442 (2013)
20. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the ICML, pp. 689–696 (2011)
21. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: Proceedings of the IJCAI, pp. 3846–3853 (2016)
22. Peng, Y., Qi, J., Huang, X., Yuan, Y.: CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network. IEEE TMM **20**(2), 405–420 (2017)
23. Peng, Y., Qi, J., Yuan, Y.: CM-GANs: cross-modal generative adversarial networks for common representation learning. arXiv preprint arxiv:1710.05106 (2017)
24. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE TPAMI **36**(3), 521–535 (2014)
25. Quadrianto, N., Lampert, C.H.: Learning multi-view neighborhood preserving projections. In: Proceedings of the ICML, pp. 425–432 (2011)
26. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's mechanical turk. In: NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147 (2010)
27. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the ACM MM, pp. 251–260 (2010)
28. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proceedings of the ICML, pp. 1060–1069 (2016)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arxiv:1409.1556 (2014)
30. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. **2**(1), 207–218 (2014)
31. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines. In: Proceedings of the NIPS, pp. 2639–2664 (2012)
32. Su, J., Zeng, J., Xiong, D., Liu, Y., Wang, M., Xie, J.: A hierarchy-to-sequence attentional neural machine translation model. IEEE TASLP **26**(3), 623–632 (2018)
33. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the CVPR, pp. 2962–2971 (2017)
34. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the ACM MM, pp. 154–162 (2017)
35. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE TPAMI **38**(10), 2010–2023 (2016)
36. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: Proceedings of the ICCV, pp. 2088–2095 (2013)
37. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv preprint arxiv:1607.06215 (2016)

38. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. VLDBJ **25**(1), 79–101 (2016)
39. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
40. Yuan, Z., Sang, J., Liu, Y., Xu, C.: Latent feature learning in social media network. In: Proceedings of the ACM MM, pp. 253–263 (2013)
41. Zhai, D., Chang, H., Shan, S., Chen, X., Gao, W.: Multiview metric learning with global consistency and local smoothness. ACM Trans. Intell. Syst. Technol. **3**(3), 53:1–53:22 (2012)
42. Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE TCSVT **24**(6), 965–978 (2014)
43. Zhai, X., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Proceedings of the AAAI, pp. 1198–1204 (2013)
44. Zhu, L., Chen, Y., Ghamisi, P., Benediktsson, J.A.: Generative adversarial networks for hyperspectral image classification. IEEE TGARS **56**(9), 5046–5063 (2018)