UNSUPERVISED CROSS-MODAL RETRIEVAL THROUGH ADVERSARIAL LEARNING

Li He[1*] *Xing Xu*[2*] *Huimin Lu*[3] *Yang Yang*[2] *Fumin Shen*[2] *Heng Tao Shen*[2]

[1]Qualcomm Technologies, Inc., San Diego, CA
[2]University of Electronic Science and Technology of China, China
[3]Kyushu Institute of Technology, Japan

ABSTRACT

The core of existing cross-modal retrieval approaches is to close the gap between different modalities either by finding a maximally correlated subspace or by jointly learning a set of dictionaries. However, the statistical characteristics of the transformed features were never considered. Inspired by recent advances in adversarial learning and domain adaptation, we propose a novel **U**nsupervised **C**ross-modal retrieval method based on **A**dversarial **L**earning, namely UCAL. In addition to maximizing the correlations between modalities, we add an additional regularization by introducing adversarial learning. In particular, we introduce a modality classifier to predict the modality of a transformed feature. This can be viewed as a regularization on the statistical aspect of the feature transforms, which ensures that the transformed features are also statistically indistinguishable. Experiments on popular multimodal datasets show that UCAL achieves competitive performance compared to state of the art supervised cross-modal retrieval methods.

Index Terms— Cross-modal retrieval, adversarial learning

1. INTRODUCTION

Cross-modal retrieval has become more important than ever with the explosion of online multimedia data. The challenge of cross-modal retrieval lies within the fact that features of different modalities have very different statistical characteristics, rendering it impossible to directly compare features of different modalities. Current research has been focused on two aspects: correlation maximization and feature selection [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]. Subspace learning and dictionary learning are popular approaches. With subspace learning, a common subspace and corresponding transforms are learned so that the transformed features are maximally correlated [1]. With dictionary learning, multiple dictionaries are joint learned by correlating the sparse coefficients obtained on the training data [3]. Mixed norm regularization has been added to improve feature selection [2] [4] [6] [7]. These methods achieve considerable performance; however, most of them are supervised and require labeled data, which could be hard to obtain in real world.

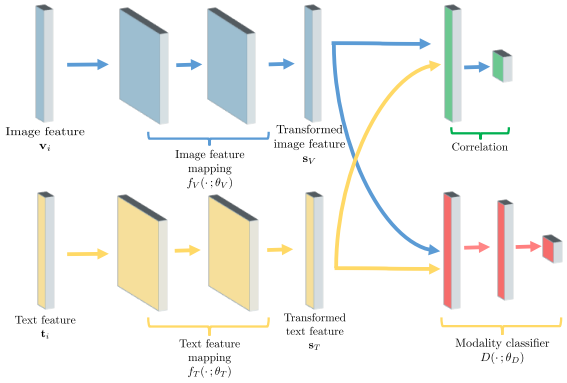


Fig. 1. The general flowchart of the proposed UCAL for cross-modal retrieval.

In the deep learning realm, several unsupervised models based on canonical correlation analysis (CCA) [11] or autoencoder have been proposed to learn modality invariant features [12] [13] [14] [15] without supervising labels. Similarly, these models learn transforms that maximize the correlation of transformed features under certain measure in the common subspace. These transforms are expected to be modality invariant so that the transformed features have similar statistical characteristics and cannot be distinguished from each other. However, existing approaches fail to explicitly address the statistical aspect of the transformed features, hence these features can still be statistically different. To enforce statistical similarity between transformed features of different modalities, similarity between their distributions must measured in certain way. However, it is never a trivial task to measure the similarity between two unknown yet complex distributions.

In order to address the statistical aspect of the feature transforms, we propose a novel approach, termed unsupervised cross-modal retrieval with adversarial learning (UCAL). UCAL is inspired by recent advance in domain adaptation [16] where adversarial learning is utilized to avoid domain shift and to facilitate generation of domain invariant features. Fig. 1 illustrates the general framework of UCAL, consisting of four major components: image feature mapping (blue), text feature mapping (yellow), modality classifier (red)

∗ Equal contributions.

and feature correlation (green). Similar to [13] [14] [15], in UCAL, we adopt two feedforward networks as the image and text feature mappings to transform the respective features to a common subspace so that they have the same dimension. In addition to requiring the transformed features to be maximally correlated, we also require them to be statistically indistinguishable in the subspace. To achieve this, we introduce modality classifier to identify the source modality of a transformed feature. These components are trained under the adversarial learning framework. This is quite different from previous methods where no requirement is placed on the statistical characteristics of the transformed features. The adversary introduced by the modality classifier can be seen as a regularization term in the subspace learning procedure of the proposed method. Therefore, it ensures that the transformed features of different modalities can be directly compared in the subspace with their intrinsic characteristics are well preserved.

Our contribution can be summarized as: 1) We introduce the adversarial learning framework into cross-modal retrieval as regularization and propose a novel unsupervised cross-modal retrieval approach that directly addresses the statistical aspect of feature transforms. 2) We develop an efficient parameter learning scheme to optimize the proposed model directly through stochastic gradient descent without additional layers. 3) We evaluate the proposed approach on four multimodal benchmark datasets and show that the our framework outperforms or achieves competitive performance compared to the state of the arts supervised learning methods.

2. RELATED WORK

2.1. Cross-modal Retrieval

Cross-modal retrieval methods have been following two main lines: correlation maximization and feature selection.

Subspace learning methods, such as Canonical Correlation Analysis (CCA) [11], have been used in or extended for cross-modal retrieval. By assuming that the representations in different features spaces are correlated through certain common information, Rasiwasia *et al*. [1] proposed to learn the subspace by maximizing the correlation between the image feature and the text feature spaces through CCA.

Dictionary learning has been introduced to address the fact that the subspace assumption could be restrictive for some real world multimodal data. Zhuang *et al*. [3] extends unimodal dictionary learning framework to multimodal data. Instead of independently learning the dictionary and corresponding coefficients for a single modality, the coefficients for different modalities are correlated using a linear mapping; $l_{1,2}$ norm was also used to discover inter-modality structures.

As pointed out by Gu *et al*. [2], both subspace and dictionary learning have problem with feature selection: either all features are linearly combined or only some components are selected from a feature vector. To tackle this, they formulated subspace learning using graph embedding and applied $l_{2,1}$ regularization to jointly perform feature selection and subspace learning. Wang *et al*. [4] proposed to explicitly learn two projections that map two modalities into a coupled common subspace and adopted $l_{2,1}$ norm on the learned projections to perform feature selection. Xu *et al*. [6] [7] further introduced dictionary learning into the coupled feature mapping framework, forming a two step framework. In particular, two dictionaries were learned jointly in a way similar to [3]; then the learned sparse representations were then mapped into a common subspace.

Meanwhile, neural networks have also been applied to cross-modal retrieval. Srivastava *et al*. [12] applied autoencoder and Restricted Boltzman Machine (RBM) to multimodal data. They followed similar pattern by adding a shared representation layer to correlate each modality. Another autoencoder based model is Correspondence Autoencoder (Corr-AE) [13]. Instead of reconstructing via shared representations, Corr-AE correlates representations learned by each autoencoder through a predefined similarity measure. The model is trained to minimize the reconstruction error for each modality and the pairwise discrepency between the learned representations. In addition, CCA has also been extended to deep CCA where neural networks were used to transform features and correlation was maximized over the whole training data [14] [15].

A major drawback falls in the statistical aspect. We know that features of different modality have different statistical properties. Although these methods tried to maximally correlate different modalities and to better choose features, none of them explicitly address the statistical aspect of the representations learned from different modalities. The transformed features are not guaranteed to possess similar statistical properties, which can make them statistically separate. In this paper, we explicitly address this issue through adversarial learning.

2.2. Adversarial Learning

Adversial learning was recently proposed by Goodfellow *et al*. [17] in GAN for image generation. Despite its extensive application in image generation [17] [18], researchers also uses it as a regularizer [16]. Ganin *et al*. [16] regularized feature extractor in domain adaptation with adversarial network to generate domain invariant features and achieved exciting performance. Yet, no attempt has been made to apply adversarial learning to cross-modal retrieval.

Inspired by these works, we introduce adversarial learning as regularization into cross-modal retrieval for image and text. Similar to the neural networks based methods, we use neural networks for feature transforms. However, we not only maximize the correlation between the transformed features, we also regularize their distributions through the introduction of modality classifier, which predicts the source modality of a transformed feature and thus brings adversary.

3. PROPOSED FRAMEWORK

3.1. Preliminaries

The original adversarial learning framework [17], *i.e*. Generative Adversarial Nets (GAN), consists of two key parts: the generator $G(\mathbf{x};\theta_G)$ and the discriminator $D(\mathbf{x};\theta_D)$, each of which is a feedforward net governed by respective parameters θ_G and θ_D. Let $p_{\mathbf{x}}(\mathbf{x})$ be the distribution of data $\mathbf{x}$. The generator G maps a random vector $\mathbf{z}$ from a prior distribution $p_{\mathbf{z}}(\mathbf{z})$ into the data space. The discriminator D outputs a scalar indicating if $\mathbf{x}$ comes from the data distribution $p_{\mathbf{x}}(\mathbf{x})$. D is trained to maximize its prediction precision while G is trained to minimize the prediction precision of D.

Goodfellow *et al*. [17] showed that for an ideal discriminator, the overall loss of the model is equivalent to the Jensen-Shannon divergence between the generator distribution p_G and $p_{\mathbf{x}}$. By properly training the generator and the discriminator, GAN is able to approximate $p_{\mathbf{x}}(\mathbf{x})$ with $p_G(\mathbf{x})$. This framework provides a means to close the gap between two different distributions. In cross-modal retrieval, this enables us to put an explicit requirement of statistical properties on the feature mappings and the learned subspace.

3.2. Adversarial Cross-modal Retrieval

Problem formulation. Let $\mathcal{D} = \{I_1, ..., I_n\}$ be a collection of n instances with each instance $I_i = (\mathbf{v}_i, \mathbf{t}_i)$ consisting of d_{V} dimensional visual feature $\mathbf{v}_i$ and d_{T} dimensional text feature $\mathbf{t}_i$. We also define feature matrices of two modalities as $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_n\}$. It is expected that the visual features and the text features have very different statistical properties and follow unknown yet complex distributions; therefore they cannot be directly compared against each other for cross-modal retrieval.

Suppose we have two mappings $f_V(\mathbf{v};\theta_V)$ and $f_T(\mathbf{t};\theta_T)$ that respectively transform the visual and text features into d dimensional vectors $\mathbf{s}_V$ and $\mathbf{s}_T$,

$$\mathbf{s}_V^i = f_V(\mathbf{v}_i;\theta_V), \quad (1)$$

$$\mathbf{s}_T^i = f_T(\mathbf{t}_i;\theta_T). \quad (2)$$

Since the transformed features are expected to be modality invariant, we drop the modality indicators V or T when we do not care the source of a transformed feature. For convince we also define $\mathbf{S}_V = \{\mathbf{s}_V^1, ..., \mathbf{s}_V^n\}$, $\mathbf{S}_T = \{\mathbf{s}_T^1, ..., \mathbf{s}_T^n\}$ and $\mathbf{S} = \{\mathbf{s}_V, \mathbf{s}_T\}$. We assume that f_V and f_T can take $\mathbf{V}$ and $\mathbf{T}$,

$$\mathbf{S}_V = f_V(\mathbf{V};\theta_V), \quad (3)$$

$$\mathbf{S}_T = f_T(\mathbf{T};\theta_T). \quad (4)$$

The transformed features are not guaranteed to be directly comparable through these transforms since they may still carry the statistical properties of their sources. Yet, existing methods, either based on subspace learning, dictionary learning or deep neural networks, focus on maximizing the correlation in the transformed space or choosing better features. No explicit requirements are imposed on the statistical aspect.

To make the features directly comparable, we require that 1) $\mathbf{s}_V$ and $\mathbf{s}_T$ must be strongly correlated and 2) the distributions of $\mathbf{s}_V$ and $\mathbf{s}_T$ must be close to each other. We use feedforward networks for feature mappings and train the whole model using the adversarial learning framework. This allows us to put an additional restriction on the statistical properties on the transformed features.

The proposed model. The proposed model is shown in Fig.1. For simplicity, we assume that the features have already been extracted from images and text. Image and text features first pass through respective transforms f_V and f_T. These transforms are inspired by subspace learning based methods where the original features are usually mapped into a common space.

To enforce the aforementioned statistical requirement, we introduce the modality classifier $D(\mathbf{s};\theta_D)$. It is equivalent to the discriminator in GAN. The classifier takes a transformed feature vector $\mathbf{s}$ and outputs a scalar indicating if $\mathbf{s}$ is from a visual or text feature. The feature transforms and the classifier are trained under two very different criteria under the adversarial learning framework. For the classifier, the goal is to maximize its prediction precision given a transformed feature vector. On the contrary, the feature transforms are trained to generate modality invariant features minimizing the classifier's prediction precision.

Formally, let $L_D(\mathbf{s}^i)$ be the modality classification loss and $L_C(\mathbf{S}_V, \mathbf{S}_T)$ be the correlation loss (*i.e*. dissimilarity under certain measure) of the transformed features. At training time, L_D can be easily determined since the source of $\mathbf{s}^i$ is known. The modality classifier losses can thus be written as $L_D(f_V(\mathbf{v}_i;\theta_V))$ and $L_D(f_T(\mathbf{t}_i;\theta_T))$ for I_i using Eq.1 and Eq.2. The correlation loss can be written as $L_C(f_V(\mathbf{V};\theta_V), f_T(\mathbf{T};\theta_T))$ using Eq.3 and Eq.4. This yields to the overall loss for the model as

$$\begin{aligned} &L(\theta_V, \theta_T, \theta_D) = L_C(f_V(\mathbf{V};\theta_V), f_T(\mathbf{T};\theta_T)) - \\ &\sum_i (L_D(D(f_V(\mathbf{v}_i;\theta_V);\theta_D)) + L_D(D(f_T(\mathbf{t}_i;\theta_T);\theta_D))) \\ &= L_C(f_V(\mathbf{V};\theta_V), f_T(\mathbf{T};\theta_T)) - \sum_i L_D^i(\theta_V, \theta_T, \theta_D), \end{aligned} \quad (5)$$

where L_D^i denote the per-sample loss and we combined visual and text features to make the notation compact.

Suppose the optimal parameters are $\hat{\theta}_V$, $\hat{\theta}_T$ and $\hat{\theta}_D$, then we have

$$\hat{\theta}_V, \hat{\theta}_T = \arg\min L(\theta_V, \theta_T, \hat{\theta}_D), \quad (6)$$

$$\hat{\theta}_D = \arg\max L(\hat{\theta}_V, \hat{\theta}_T, \theta_D). \quad (7)$$

This follows the adversarial learning framework: the feature transforms are trained to generate modality invariant features to maximize the classifier error while the classifier is trained to minimize its error. If the modality classifier were not present, the framework would simply try to maximize the correlation, falling back to a CCA-style model. The adversary

introduced by the modality classifier can be considered as a regularizer for the transformed distributions.

Correlation loss. Optimizing the correlation loss L_C over the entire data is similar to CCA and is complicated in the context of neural networks. As pointed in [11], the original CCA objective can be viewed as minimizing the distance between the learned projections of the two views while satisfying the whitening constraints for the projections and this constraint complicates the optimization. Inspired by [13] and [19], we use pairwise l_2 norm as the correlation loss. Specifically,

$$L_C(f_V(\mathbf{V};\theta_V), f_T(\mathbf{T};\theta_T)) = \sum_i \|f_V(\mathbf{v}_i;\theta_V) - f_T(\mathbf{t}_i;\theta_T)\|_2 . \tag{8}$$

Eq.8 explicitly imposes requirement on the distributions of the transformed features, which is expected to give stronger promise on preserving the statistical properties of different modalities. Meanwhile, it is different from the regularization term in Corr-AE that accounts for the reconstruction errors of originally multi-modal features.

Optimization. The loss function in Eq. 5 cannot be directly optimized using SGD [16]. The difficulty comes with the opposite optimization goals for the feature transforms and the classifier. Instead of using a reversal layer as in [16], we reformulate the loss. To unify the training goal, we proposed to compare against incorrect predictions rather than the correct ones, in which case the classification loss is to be minimized. This removes the difficulty and allows us to use SGD directly. Denoting the new loss as $L_{D'}$, we have $L_D = -L_{D'}$.

To simultaneously train the three components using Eq. 5 we limit the update to respective components. In particular, the modality classifier loss can be further divided into: $L^V_{D'}(\theta_V, \theta_D)$ and $L^T_{D'}(\theta_T, \theta_D)$, representing the classification loss for visual and text modality, respectively. Note that the reformulated loss is used here. We rewrite Eq. 5 as

$$L(\theta_V, \theta_T, \theta_D) = L_C(\theta_V, \theta_T) + L^V_D(\theta_V, \theta_D) + L^T_D(\theta_T, \theta_D). \tag{9}$$

With Eq. 9, we update the model parameters using the following rules

$$\theta_V \leftarrow \theta_V - \mu \frac{\partial}{\theta_V}(L_C + L^V_{D'}), \tag{10}$$

$$\theta_T \leftarrow \theta_T - \mu \frac{\partial}{\theta_T}(L_C + L^T_{D'}), \tag{11}$$

$$\theta_D \leftarrow \theta_D - \mu \frac{\partial}{\theta_D}(L^V_{D'} + L^T_{D'}). \tag{12}$$

where μ is the learning rate (which can vary over time).

4. EXPERIMENTS

4.1. Experimental Setup

Datasets. We evaluate our framework using four publicly available datasets: Wiki [20], Pascal Sentences [21], Flickr30K [22] and MS COCO [23]. Each instance in the four datasets contains features from both image and text modalities, as well a single or multiple labels. Table 1 summarizes the general statistics of these datasets.

Table 1. General information of the four benchmark datasets. Note that the column of "Instances" is formatted as number of training and test instances; "(S)" and "(D)" stand for shallow and deep feature representation, respectively. Besides, "-" means no feature representation is used.

Dataset	Instances	Labels	Image feature	Text feature
Wiki	1300, 1566	10	(S) 128d SIFT (D) 4096d VGG	10d LDA 300d Word2Vec
Pascal Sentence	800, 200	20	(S) 512d Gist (D) 4096d VGG	250d word frequency 300d Word2Vec
Flickr30K	28917, 2866	350	(S) 512d Gist (D) 4096d VGG	3000d word frequency 300d Word2Vec
MS COCO	66226, 16557	500	(S) - (D) 4096d VGG	- 300d Word2Vec

Features. We adopt shallow and deep feature representations for both image and text modalities. The shallow feature representations for images and texts are similar to [1][4][3], where low-level visual descriptions and word frequency are used to represent images and texts, respectively. Because of the recent progress in representing images and texts using deep neural network, we also extract deep feature representations. Specifically, we use the 4096d output vector of *fc2* layer of pretrained VGG-16 network [24] to represent each image. For each text, we first use publicly available GloVe [25] trained on Wikipedia 2014 with 6B tokens to generate a 300d vector for each word in the text, and then calculate the mean vector for all the words to represent the text.

Network architecture. We set the dimension of the transformed features to 200. We use a three layer network $4096 \rightarrow 2048 \rightarrow 1024 \rightarrow 200$ for image feature transform and a single layer network $300 \rightarrow 200$ for text feature transform. For the modality classifier, we use a three layer network $200 \rightarrow 100 \rightarrow 50 \rightarrow 2$. We use binomial cross-entropy for loss functions L_D. While training our model we notice that a strong modality classifier on the contrary can worsen the performance. To alleviate this, we update the modality classifiers less often then the feature transforms.

Baselines. We compare our proposed method against CCA [1], Bi-DBN [12], Corr-AE [13], SliM [3], LCFS [4], CCA-3V [26], CDLFM [6]. The baselines are chosen from state of the art methods based on subspace or dictionary learning and deep neural networks. CCA, Bimodal DBN and Corr-AE are unsupervised methods, where CCA is a classical method based on subspace learning, Bi-DBN and Corr-AE are based on neural networks. LCFS and CDLFM are supervised methods based on subspace learning and dictionary learning respectively. Both methods utilize labels as supervision and performs feature selection. We compare against these two methods to see if our novel regularization can compensate the absence of supervision and feature selection to certain extent.

Evaluation metrics. We apply UCAL to two cross-modal

Table 2. Cross-modal retrieval comparison in terms of mAP on four datasets. Here ‡ means the numbers of the method are cited from the original paper, the other numbers are obtained from our implementation. "-" means no repeated result available yet. The best results on each dataset are highlighted in bold font.

Methods	Wiki			Pascal Sentence			Flickr30K			MSCOCO		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA	0.254‡	0.184‡	0.219	0.263	0.219	0.241	0.228	0.245	0.237	-	-	-
Bi-DBN	0.189‡	0.222‡	0.205	-	-	-	-	-	-	-	-	-
Corr-AE	0.276‡	0.234‡	0.255	-	-	-	-	-	-	-	-	-
SliM	0.255‡	0.202‡	0.228	-	-	-	-	-	-	-	-	-
CCA-3V	0.275	0.224	0.249	0.316	0.270	0.293	0.294	0.312	0.303	-	-	-
LCFS	0.279‡	0.214‡	0.246	0.331	0.265	0.298	0.271	0.296	0.284	-	-	-
CDLFM	0.272‡	0.223‡	0.247	0.327	0.281	0.304	0.293	0.301	0.297	-	-	-
CCA	0.267	0.222	0.244	0.349	0.292	0.321	0.285	0.304	0.295	0.122	0.245	0.184
CCA-3V	0.284	0.249	0.266	0.352	0.297	0.325	0.363	0.396	0.380	0.177	0.315	0.246
LCFS	**0.296**	0.254	**0.275**	0.442	**0.357**	**0.399**	0.387	0.415	0.401	0.183	**0.342**	0.263
CDLFM	0.283	0.264	0.273	0.432	0.322	0.377	0.392	**0.431**	**0.412**	0.179	0.326	0.253
Proposed	0.263	**0.273**	0.268	**0.448**	0.325	0.387	**0.397**	0.422	0.409	**0.189**	0.338	**0.264**

retrieval tasks, *i.e*. image retrieval by text (Img2Txt) and text retrieval by image (Txt2Img). To evaluate the performance, we use the standard measure of mean average precision (mAP) and precision-scope curve that have been widely adopted in literatures [1][14][4][13]. The experiments are conducted on a desktop machine with a 4-core CPU at 4GHz and 32 GB memory.

4.2. Experimental Results

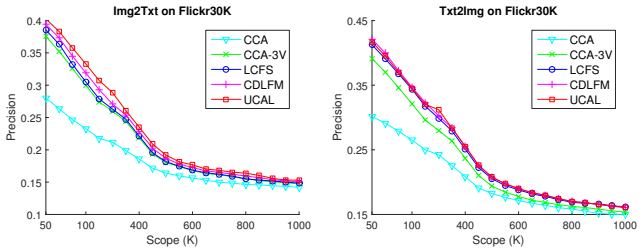
Overall Comparison with Baselines. For our UCAL and baselines, we report their cross-modal retrieval performance on all datasets in Table 2 with shallow and deep feature representations for both image and text modalities (shallow features in the upper and deep features in the lower part).

We see that on small datasets Wiki and Pascal Sentence, the retrieval performance of CCA, CCA-3V, LCFS and CDLFM with deep feature representations are much better than those those using shallow ones, indicating that deep representations of multi-modal data are beneficial for better retrieval performance. In addition, the unsupervised learning methods (CCA, Bi-DBN and Corr-AE) performs inferior to the supervised ones (CCA-3V, LCFS and CDLFM). The reason is that the unsupervised methods are limited to pair-wise closeness in the common subspace, while the supervised ones also utilize additional class information to obtain better separation between classes in the common subspace. Regarding the proposed UCAL, it outperforms the unsupervised CCA (except for Img2Txt on Wiki) and is competitively compared to the supervised methods. This indicates that our strategy is actually effective in closing the modality gaps. The supervised methods LCFS and CDLFM are able to obtain highest scores. This is understandable because they also utilize label information and perform feature selection to further enhance the performance. Corr-AE outperforms our method by a small margin in image to text task; however, our method achieves better average performance. Specifically, it is worth noting that the text on Wiki dataset appears as paragraphs rather than short sentences, posing some difficulty. However, the proposed UCAL still achieves good performance on Text2Img task, showing its advantage in obtaining more effective transformed features for the text modality.

On larger datasets Flickr30K and MS COCO, the proposed UCAL is compared with the baselines CCA, CCA-3V, LCFS and CDLFM with deep feature representations. For the four baselines, Principle Component Analysis (PCA) is first performed on the original deep feature representations for image modality to remove redundancy. On these two datasets, the text mainly appears as short sentences. Thus the text feature representation generated by UCAL consistently makes more sense in this situation, resulting in stronger retrieval performance on Text2Img than that on Img2Txt. It is in fact exciting to see that our method achieves competitive performance compared to LCFS and CDLFM on average, both of which are supervised learning methods that utilize labels as supervision and mix norm regularization for feature selection. This indicates that the adversarial regularization is working as expected.

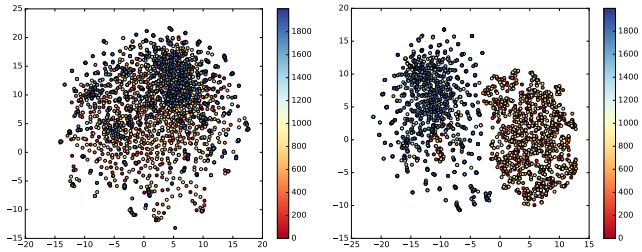
Detailed Analysis on UCAL. To further investigate the effectiveness of the proposed UCAL on both Img2Txt and Txt2Img tasks, we plot the precision-scope curves of UCAL and several baselines in Fig. 2 with various scopes (*i.e*., the top k retrieved instances). We observe that UCAL outperforms its counterparts on Img2Txt task and performs on par with the supervised baselines LCFS and CDLFM on Txt2Img task. Thus, it again validates the advantage of UCAL on modeling cross-modal data for retrieval tasks.

To further understand the effect of the adversary loss in our UCAL framework, we visualize the transformed features using t-SNE [27] on Flickr30K dataset. In particular, for each of the image and text modality we randomly choose 1000 transformed features in the test set to form a total of 2000 features. The chosen features do not necessarily form image text pairs. We then use t-SNE to visualize the distribution of these features. The intuition of introducing the adversary is to close the statistical gaps. Therefore, the transformed features are expected to form a single cluster. Fig. 3 shows the t-SNE embedding for the test data of Flickr30k dataset. Fig. 3(a) shows the features with adversarial loss and Fig. 3(b) shows



(a) Img2Txt (b) Txt2Img

Fig. 2. Performance of different methods on Flickr30K dataset, based on precision-scope curve for $K = 50$ to 1000.



(a) With adversarial loss (b) Without adversarial loss

Fig. 3. t-SNE visualization for the test data in Flickr30k. The red colors represent the visual features and the blue colors represent text features.

the same without adversarial loss. We can see that without adversarial loss, the transformed features are still scattered and the adversarial loss indeed effectively closes the gap between different modalities. This indicates that adversarial learning as a regularization works as expected to close the statistical gaps between modalities and that it is an effective tool for processing multimodal data.

5. CONCLUSION

In this paper we proposed UCAL, a novel unsupervised model for cross-modal retrieval utilizing the adversarial learning framework. In addition to exploiting the correlation between mutlimodal features, UCAL also explicitly accounts for the statistical properties of the transformed features through the modality classifier and the adversary against feature transforms. Although our model does not use any supervision information, it is able to outperform or achieve competitive performance compared to many state of the art methods. We believe adversarial learning can become a powerful tool in cross-modal data analysis.

Acknowledgements. X. Xu *et al.* [2] were supported by NSFC grants 61502081, 61572108, 61632007, 61472063; China Thousand-Young-Talents Program; Fundamental Research Funds for Central Universities ZYGX2014Z007, ZYGX2015J055 and ZYGX2016KYQD114. H. Lu[3] was supported by LEADER of MEXT-Japan (16809746), The Telecommunications Foundation, REDAS and SCAT.

6. REFERENCES

[1] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*. ACM, 2010, pp. 251–260.

[2] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *IJCAI*, 2011.

[3] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *AAAI*, 2013.

[4] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *ICCV*, 2013, pp. 2088–2095.

[5] Yang Yang, Zheng-Jun Zha, Yue Gao, Xiaofeng Zhu, and Tat-Seng Chua, "Exploiting web images for semantic video indexing via robust sample-specific loss," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1677–1689, 2014.

[6] X. Xu, A. Shimada, R. Taniguchi, and L. He, "Coupled dictionary learning and feature mapping for cross-modal retrieval," in *ICME*. IEEE, 2015, pp. 1–6.

[7] X. Xu, Y. Yang, A. Shimada, R. Taniguchi, and L. He, "Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts," in *ACM MM*. ACM, 2015, pp. 847–850.

[8] Yang Yang, Hanwang Zhang, Mingxing Zhang, Fumin Shen, and Xuelong Li, "Visual coding in a semantic hierarchy," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 59–68.

[9] Xing Xu, Li He, Atsushi Shimada, Rin-ichiro Taniguchi, and Huimin Lu, "Learning unified binary codes for cross-modal retrieval via latent semantic hashing," *Neurocomputing*, vol. 213, pp. 191–203, 2016.

[10] Lingyang Chu, Yanyan Zhang, Guorong Li, Shuhui Wang, Weigang Zhang, and Qingming Huang, "Effective multimodality fusion framework for cross-media topic detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 556–569, 2016.

[11] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[12] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NIPS*, 2012, pp. 2222–2230.

[13] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM MM*. ACM, 2014, pp. 7–16.

[14] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis.," in *ICML*, 2013, pp. 1247–1255.

[15] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015, pp. 3441–3450.

[16] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *NIPS*, 2014, pp. 2672–2680.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015, pp. 1083–1092.

[20] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *TPAMI*, vol. 36, no. 3, pp. 521–535, March 2014.

[21] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[22] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015, pp. 2641–2649.

[23] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[26] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *IJCV*, vol. 106, no. 2, pp. 210–233, 2014.

[27] L. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.