World Wide Web
https://doi.org/10.1007/s11280-018-0541-x

Deep adversarial metric learning for cross-modal retrieval

Xing Xu[1] · Li He[3] · Huimin Lu[4] · Lianli Gao[1] · Yanli Ji[2]

Received: 15 August 2017 / Revised: 18 February 2018 / Accepted: 27 February 2018

Abstract Cross-modal retrieval has become a highlighted research topic, to provide flexible retrieval experience across multimedia data such as image, video, text and audio. The core of existing cross-modal retrieval approaches is to narrow down the gap between different modalities either by finding a maximally correlated embedding space. Recently, researchers leverage Deep Neural Network (DNN) to learn nonlinear transformations for each modality to obtain transformed features in a common subspace where cross-modal matching can be performed. However, the statistical characteristics of the original features for each modality

This article belongs to the Topical Collection: *Special Issue on Deep vs. Shallow: Learning for Emerging Web-scale Data Computing and Applications*
Guest Editors: Jingkuan Song, Shuqiang Jiang, Elisa Ricci, and Zi Huang

✉ Huimin Lu
dr.huimin.lu@ieee.org

Xing Xu
xing.xu@uestc.edu.cn

Li He
lih@qti.qualcomm.com

Lianli Gao
lianli.gao@uestc.edu.cn

Yanli Ji
yanliji@uestc.edu.cn

[1] Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

[2] School of Automation, University of Electronic Science and Technology of China, Chengdu, China

[3] Qualcomm Technologies, Inc., San Diego, CA, USA

[4] Kyushu Institute of Technology, Fukuoka, Japan

Published online: 04 April 2018

are not explicitly preserved in the learned subspace. Inspired by recent advances in adversarial learning, we propose a novel Deep Adversarial Metric Learning approach, termed DAML for cross-modal retrieval. DAML nonlinearly maps labeled data pairs of different modalities into a shared latent feature subspace, under which the intra-class variation is minimized and the inter-class variation is maximized, and the difference of each data pair captured from two modalities of the same class is minimized, respectively. 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, which ensures that the transformed features are also statistically indistinguishable. Experiments on three popular multimodal datasets show that DAML achieves superior performance compared to several state of the art cross-modal retrieval methods.

Keywords Cross-modal retrieval · Adversarial learning · Metric learning

1 Introduction

Over the past few years, multi-modal data, i.e. media data of various types but homogeneous topic, has been growing rapidly with the emerging development of social media websites (e.g., Twitter, Facebook, Youtube, Instagram, etc), where users are allowed to retrieve information from these heterogeneous data using their preferred queries [22, 26, 28, 29, 49]. In order to maximally benefit from the richness of multimedia data and make optimal use of the rapidly developing multimedia technology, automated mechanisms are needed to establish a similarity link from one multimedia item to another if they are related to each other, independent of the type of modalities, such as text, visual or audio, present in the items. In order to provide an answer to the above challenge, research towards reliable solutions for cross-modal retrieval, that are able to operate across modality boundaries, has gained significant attraction recently.

The primary issue in cross-modal retrieval lies within the fact that features of different modalities have very different statistical characteristics, indicating its impossibility to directly compare features of different modalities. Current research has been focused on two aspects: correlation maximization [9, 22, 48, 49] and feature selection [35, 40, 42, 43, 45, 46] [2]. 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 [22]. With dictionary learning, multiple dictionaries are jointly learned by correlating the sparse coefficients obtained on the training data [49]. Mixed norm regularization has been added to improve feature selection [9, 35, 42, 43]. These methods achieve considerable performance; however, most of them are supervised and require labeled data, which could be hard to obtain in the real world.

In the deep learning realm, several unsupervised models based on canonical correlation analysis (CCA) [10] or autoencoder have been proposed to learn modality invariant features [1, 5, 31, 44] without supervising labels. These models generate representations in an embedding space shared by different modalities and optimizations are performed to maximize the correlation for the shared representation. The core of these approaches is to close the gap between different modalities by finding certain transforms under which the transformed features are maximally correlated. 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, especially the intra-modal discriminativeness and the inter-modal consistence, hence these features can still be statistically different.

In order to address the statistical aspect of the feature transforms, we propose a novel DNN based approach, termed Deep Adversarial Metric Learning (DAML), for cross-modal retrieval task. DAML is inspired by the recent advance in domain adaptation [6] where adversarial learning is utilized to avoid domain shift and to facilitate generation of domain invariant features. Besides, to enforce statistical similarity between transformed features of different modalities, similarity between their distributions must be measured in a certain way. In our proposed DAML, we also employ the coupled metric learning technique [15] to learn an appropriate similarity measure that preserves the statistical similarity between transformed features of different modalities.

Figure 1 illustrates the general framework of DAML. Similar to [1, 5, 38, 44], we adopt two feed-forward networks as the image and text feature mappings in DAML to nonlinearly transform the respective features to a common subspace, under which the intra-class variation is minimized and the inter-class variation is maximized. In addition to requiring the transformed features to be maximally correlated, we also require them to be statistically indistinguishable in the subspace, i.e. the difference of each sample pair captured from two modalities of the same class is minimized. 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. By doing so, we explicitly require that mapped features of different modalities have similar statistical distributions. The adversary introduced by the modality classifier can be seen as a

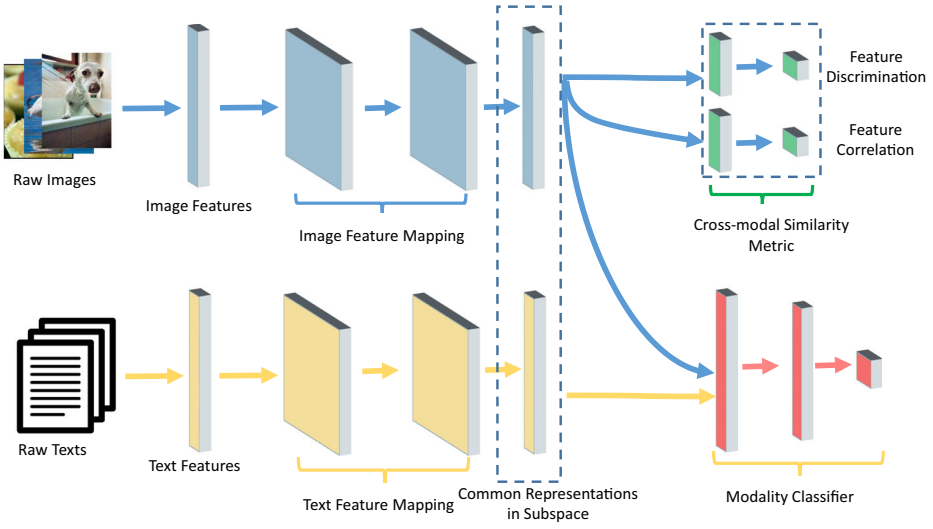


Figure 1 The general architecture of the proposed DAML consists of four major components: image feature projection, text feature projection, modality classifier, cross-modal similarity metric, which together form a standard feed-forward architecture. The image and text features are mapped to the common subspace with successive two-fold procedure. One branch termed cross-modal similarity metric proceeds the feature discrimination and feature correlation jointly in the subspace, and the other branch termed modality classifier accounts for the diversity between the representations of different modalities in the subspace. Adversarial learning manner is adopted to jointly optimized the two branches during training

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.

This paper is an extension and improvement of our previous method termed UCAL presented in [11]. The main differences between the proposed DAML and previous UCAL can be summarized as the following three aspects: 1) our proposed DAML is a supervised cross-modal learning approach that incrementally incorporates the discriminativeness of class labels in the learned transformed features, while UCAL is an unsupervised method that limitedly maximizing the correlation of inter-modal data; 2) our proposed DAML also employs coupled metric learning technique to learn appropriate distance metric that preserve the statistical distribution of multimodal data; 3) the parameter learning algorithm that learns the optimal neural network weights developed for the proposed DAML is also different from that in UCAL, since the weights play the roles of both transformations and distance metric in DAML. Comprehensive evaluation on three benchmark datasets illustrates that our proposed DAML significantly outperforms previous UCAL and several other state of the art cross-modal retrieval approaches.

The rest of paper is organized as follows. In Section 2, we discusses previous work in cross-modal retrieval and adversarial learning. We describe details of the proposed method in Section 3 and present the experimental results in Section 4. Finally, the conclusion is made in Section 5.

2 Related work

2.1 Cross-modal retrieval

As for the traditional cross-modal retrieval methods, one popular group is subspace learning based methods, such as Canonical Correlation Analysis (CCA) [10] and its extensions [7, 22, 23, 48]. By assuming that the representations in different features spaces are correlated through certain common information, Rasiwasia et al. [22] proposed to learn the subspace by maximizing the correlation between the image feature and the text feature spaces through CCA. Sharma et al. [23] proposed multiview extensions to CCA, LDA and Marginal Fisher Analysis (MFA), i.e. Generalized Multiview Analysis (GMA), Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA), and showed that they performed well on cross-modal retrieval problems.

It is notable that 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. [49] 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. [9], 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. Tian et al. [32] explored the intrinsic manifold structures in different modalities and developed a so-called correlation component manifold space learning method to capture the correlations residing in the heterogeneous data. Wang et al. [35] 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. [42, 43] 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 [49]; 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. [31] 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) [5]. 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 discrepancy between the learned representations. Wang et al. [37] further adopted stacked auto-encoders to form deeper non-linear embeddings for different modalities, showing the capability of learning more effective mapping functions and shared representations. Andrew et al. [1] proposed a direct extension to CCA, namely DCCA. It uses two feedforward networks to transform features of each modality and the networks are trained to maximize the correlation between the transformed features over all the data. Yan et al. [44] further proposed an end-to-end learning framework based on DCCA. 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.

Moreover, several coupled metric learning algorithms have been proposed for cross-modal matching such as Cross Modal Metric Learning (CMML) [17], Cross-Modal Similarity Learning (CMSL) [12], Coupled Marginal Fisher Analysis (CMFA) [24] and Online Asymmetric Similarity Learning (OASL) [39]. These methods only learn a pair of linear transformations to map cross-modal samples into a new common feature space, which is not effective enough to discover the nonlinear relationship of samples. Later, Liong et al. [15] proposed Deep Coupled Metric Learning (DCML), a metric learning approach that learns two sets of nonlinear transformations to map data samples into common space considering the variation of different classes. Different from DCML, our proposed DAML is based on adversarial learning, and utilizes category information adequately to preserve inter-modal and intra-modal structure simultaneously, thus ensures that the learned subspace feature representations to be both discriminative within modality and modality-invariant.

Lastly, it is worth mention that a bundle of hashing based approaches such as [27, 40, 41, 50] have been proposed for cross-modal retrieval problem. More related works can be referred to the latest literature review in [33]. These cross-modal hashing methods find linear projections to embed the heterogeneous data into a common Hamming space, where the multi-modal features are represented by low dimensional binary codes. Different from the hashing based methods, we focus on the traditional cross-modal retrieval task and aim to learn compact real-valued subspace representations rather than binary codes.

2.2 Adversarial learning

Adversial learning was recently proposed by Goodfellow et al. [8] in GAN for image generation. The framework consists of two major components, namely the *generator* and the *discriminator*. The two components have opposite training goals: the generator is trained

to generate samples that cannot be distinguished from the source by the discriminator; the discriminator is trained to correctly identify the samples that are produced by the generator. Eventually, the generator learns to duplicate the source distribution. Despite its extensive application in image generation [8, 20], researchers also uses it as a regularizer [6]. Makhzani et al. [16] introduced adversarial learning into autoencoder by regularizing the intermediate representation of the autoencoder using a prior distribution through adversarial loss. In particular, a classifier is introduced to identify if a sample is drawn directly from the prior distribution. The encoder is trained to fool the classifier so the learned representations have a similar distribution as the prior. Larsen et al. [14] combined adversarial network with Variational Autoencoder (VAE) [13]. From the perspective of VAE, the adversarial part provides an additional adversarial loss to the VAE. This can be considered as a regularized VAE. Larsen et al. [14] used an additional adversarial network to regularize an improved version of Variational Autoencoder and proved its efficiency via image reconstruction and manipulation. A closely related work is by Ganin et al. [6], where adversarial learning was applied to domain adaptation to learn domain invariant features. Ganin et al. [6] 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 method

3.1 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 $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_n\}$. In practice, the visual features and the text features are represented in different high dimensional spaces with diverse statistical properties; therefore they cannot be directly compared against each other. Suppose we have two mappings $f_V(\mathbf{v}; \theta_V) = f_V(\mathbf{v}_i; \theta_V)$ and $f_T(\mathbf{t}; \theta_T) = f_T(\mathbf{t}_i; \theta_T)$ that respectively transform the visual and text features into d dimensional vectors $\mathbf{s}_V$ and $\mathbf{s}_T$ with same dimension.

Although the transformed features have the same dimensionality, they are not guaranteed to be directly comparable since the statistical properties of the transformed features are still unknown. These transformed features can still follow unknown yet complex distributions, which prohibits effective cross-modal retrieval. Yet, existing methods, either based on subspace 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 have the following two objectives: 1) it is desirable to exploit more discriminative information from training samples; 2) it is expected to reduce the modality gap of the pairwise data from different modalities. We use

feed-forward networks to train nonlinear transformation for each modality using the adversarial learning framework. This allows us to put an additional restriction on the statistical properties on the transformed features.

3.2 Deep adversarial metric learning

As shown in Figure 1, our proposed DAML first conducts image and text feature projection to obtain the transformed representations $\mathbf{s}_V$ and $\mathbf{s}_T$, meanwhile the constraints of the intra-modal and inter-modal similarity metric and modality classifier restrain the learned subspace representations to be discriminative and modality-invariant. In the second stage, we construct a multi-task learning architecture to learning discriminative and modality-invariant subspace representations jointly. Specifically, in the following subsections, we decompose the subspace learning procedure into three loss terms: 1) *adversarial loss* was utilized to minimize the "modality gap" between two unknown distributions of representations from different modalities to promote modality-invariant; 2) *feature discrimination loss*, which models the intra-modality similarity by category information and ensures learned representations to be discriminative; 3) *feature correlation loss*, which minimize the distances among intra-class cross-modality samples and maximizes the distances among inter-class cross-modality samples.

3.2.1 Adversarial loss

To enforce the statistical requirement and close the "heterogeneity gap" demonstrated above, a modality classifier D with parameters θ_D was introduced, which acts as the "discriminator" in GAN. Mapped features from image modality are assigned with label **01**, while mapped features from text modality are assigned with label **10**. For the modality classifier, the goal is to differentiate the source modality as precise as possible given an unknown mapped feature. For the classifier implementation, we used a 3-layer feed-forward neural network with parameters θ_D (see Section 3.3 for implementation details). The adversarial loss L_{adv} can now formally be defined as:

$$L_{adv}(\theta_V, \theta_T, \theta_A) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{m_i} \cdot (\log D(\mathbf{v}_i; \theta_A) + \log(\mathbf{1} - D(\mathbf{t}_i; \theta_A))). \tag{1}$$

Essentially, L_{adv} denotes the cross-entropy loss of modality classification all instances $o_i, i = 1, ..., n$ used per iteration for training. Furthermore, $\mathbf{m_i}$ is the ground-truth modality label of each instance, expressed as one-hot vector, while $D(.; \theta_D)$ is the generated modality probability per item (image or text) of the instance o_i.

3.2.2 Feature discrimination loss

In order to ensure that the intra-modal discrimination in data is preserved after feature projection, a classifier is deployed to predict the semantic labels of the items projected in the common subspace. For this purpose, a feed-forward network activated by softmax was added on top of each subspace embedding neural network. This classifier takes the projected features of the instances o_i of coupled images and texts as training data and generates as output a probability distribution of semantic categories per item.

Suppose l_i to be groundtruth label of each representation, which is expressed as one-hot vector. And the predicted probability distribution from outputs of label classifier is described as $\hat{p_i}$. Then the intra-modality objective function can be written as follows, regardless of which modality the transformed feature representations come from.

$$L_{dis}(\theta_V, \theta_T, \theta_D) = -\frac{1}{N}\sum_{n=1}^{N}(l_i \cdot (\log \hat{p}_i(f_V(v_i)) + \log \hat{p}_i(f_T(t_i)))). \tag{2}$$

3.2.3 Feature correlation loss

For inter-modal structure, we utilized correlation loss motivated by the coupled metric learning. The loss aims to minimize the intra-class variation and maximize the inter-class variation for feature representation of different modalities. Specifically, for each pair of training samples v_i and t_j from two different modalities, we compute their square distance as $d(v_i, t_j) = \|f_V(v_i) - f_T(t_i)\|_2^2$. We expect $d(v_i, t_j)$ to be as small as possible if v_i and t_j are of the same class and as large as possible otherwise. This can be formulated as the following constraints:

$$d(v_i, t_j) \leq \xi_1, \quad if \;\; l_{v_i,t_j} = 1, \tag{3}$$

$$d(v_i, t_j) \geq \xi_2, \quad if \;\; l_{v_i,t_j} = -1, \tag{4}$$

where $l_{v_i,t_j} = 1$ indicates that v_i and t_j belong to the same class, and $l_{v_i,t_j} = -1$ otherwise, ξ_1 and ξ_2 are the small and large thresholds, respectively. We follows [15] to integrate the large margin optimization objective:

$$L_{cor}(\theta_V, \theta_T, \theta_C) = \sum_{i,j} s(1 - l_{v_i,t_i}(\theta - d(v_i, t_i))) + \sum_i \|f_V(\mathbf{v}_i) - f_T(\mathbf{t}_i)\|_2, \tag{5}$$

where $s(\cdot)$ is a generalized logistic loss function, $\xi_1 = \xi - 1$ and $\xi_2 = \xi + 1$. In (5), the second term is similar as the correlation loss term in [11] that minimizes the difference between each pair of data of the same class captured from different modalities.

3.3 Optimization

As demonstrated above, we can incorporate three loss terms in (2), (5) and (1) altogether, which can be optimized through SGD and the optimization goals of these two objective functions are opposite, which can be formally described as a **min-max game** just as shown in [8]:

$$\begin{aligned}(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_D, \hat{\theta}_C) = \arg\min_{\theta_V,\theta_T,\theta_C,\theta_D} & \alpha L_{dis}(\theta_V, \theta_T, \theta_C, \theta_D) \\ & + \beta L_{cor}(\theta_V, \theta_T, \theta_C, \theta_D) - \sigma \cdot L_{adv}(\hat{\theta}_A),\end{aligned} \tag{6}$$

$$\hat{\theta}_A = \arg\max_{\theta_A} (\alpha L_{dis}(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_C, \hat{\theta}_D) + \beta L_{cor}(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_C, \hat{\theta}_D) - \sigma \cdot L_{adv}(\theta_A)). \tag{7}$$

Here the feature discrimination loss term L_{dis} is a classifier that predicts the semantic labels of the items projected in the common subspace, thus incorporating the discriminations of labels into the common subspace; the feature correlation loss term L_{cor} aims to minimize the intra-class variation and maximize the inter-class variation for feature representation of different modalities; and the *adversarial loss* term L_{adv} is a cross-entropy loss term used

in the modality classifier, which differentiates the source modality of image or text. Parameters α and β are the weight coefficients for the feature discrimination loss term L_{dis} and feature correlation loss term L_{cor} respectively, σ is the ratio between these two loss terms and the adversarial loss L_{adv}, which controls the balance between the two branches of the feature projection and the adversary. One way to train such an architecture has been proposed in [6], which add adversarial loss L_{adv} to embedding loss L_{emb} and utilizing Gradient Reversal Layer (GRL) (as shown in Figure 1) to incorporate min-max optimization. If a Gradient Reversal layer is added before the first layer of modality classifier, the min-max optimization can be performed simultaneously, which can be summarized as the Algorithm 1.

Algorithm 1 Pseudocode of the proposed DAML.

Require: Image feature matrix $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$; Text feature matrix $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_n\}$; Label matrix $\mathbf{L} = \{l_1, ..., l_n\}$; hyperparameters: $\xi, \alpha, \beta, \sigma$;

Ensure:

1: **for** k steps **do**
2: Update parameters of feature projectors θ_V, θ_T by **descending** their stochastic gradients:
3: $\theta_V \leftarrow \theta_V - \mu \cdot \nabla_{\theta_V} \frac{1}{m}(\alpha L_{dis} + \beta L_{cor} - \sigma L_{adv})$
4: $\theta_T \leftarrow \theta_T - \mu \cdot \nabla_{\theta_T} \frac{1}{m}(\alpha L_{dis} + \beta L_{cor} - \sigma L_{adv})$
5: **end for**
6: Update parameters of modality classifier by **ascending** its stochastic gradeints through Gradient Reversal Layer:
7: $\theta_A \leftarrow \theta_A + \mu \cdot \lambda \cdot \nabla_{\theta_A} \frac{1}{m}(\alpha L_{dis} + \beta L_{cor} - \sigma L_{adv})$
8: **return** learned representations in common subspace: $f_V(\theta_V)$ and $f_T(\theta_T)$.

4 Experiments

4.1 Experimental setup

4.1.1 Datasets and features

We conduct experiments on three widely-used cross-modal datasets: Wikipedia [4], NUS-WIDE-10k [3] and Pascal Sentence [21]. For these datasets, each image-text pair is linked by a single class label and the text modality consists of discrete tags. Here we briefly introduce the three datasets adopted in the experiment.

- Wikipedia[1] is the most widely-used dataset for cross-modal retrieval task. This dataset consists of 2,866 image/text pairs of 10 categories, and is randomly divided as follows: 2,173 pairs for training, 231 pairs for validation and 462 pairs for testing.
- Pascal Sentence[2] is generated from 2008 PASCAL development kit. This dataset contains 1,000 images which are evenly categorized into 20 categories, and each image has 5 corresponding sentences which makes up one document. For each category, 40

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

[2] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Table 1 General statistics of the four datasets used in our experiments, where "*/*" in columns of "Instance" stands for the number of training/test image-text pairs

Dataset	Instances	Labels	Image feature	Text feature
Wikipedia	1,300/1,566	10	128d SIFT 4,096d VGG	10d LDA 3,000d BoW
Pascal Sentence	800/200	20	4,096d VGG	1,000d BoW
NUS-WIDE-10K	8,000/1,000	350	4,096d VGG	1,000d BoW

documents are selected for training, 5 documents for testing and 5 documents for validation.

– NUS-WIDE-10K[3] is generated from NUS- WIDE dataset. NUS-WIDE dataset consists of about 270,000 images with their tags categorized into 81 categories. While NUS-WIDE-10k dataset has totally 10,000 image/text pairs selected evenly from the 10 largest categories of NUS-WIDE dataset, which are animal, cloud, flower, food, grass, person, sky, toy, water and window. The dataset is split into three subsets: Training set with 8,000 pairs, testing set with 1,000 pairs and validation set with 1,000 pairs.

For fair and objective comparison, we exactly follow the dataset partition and feature extraction strategies of [19, 36] in the experiments. The general statistics of the four datasets are summarized in Table 1.

It is worth mention that for all datasets, we mainly use image feature extracted from deep Convolutional Neural Network (CNN) to represent an image, as the deep visual feature has shown strong ability and been widely used for image representation. Specifically, the adopted deep feature is 4,096d vector extracted by the fc7 layer of VGGNet [25] for all compared methods on all datasets. Regarding the text feature, we use the traditional bag of words (BoW) vector with TF-IDF weighting scheme to represent each text instance, and the dimension of the BoW vector in each dataset is also illustrated in Table 1. In addition, to make fair comparison with several earlier cross-modal retrieval approaches on Wikipedia dataset, we also adopt the publicly available 128d SIFT feature for image and 10d LDA feature for text representations[4]1, respectively.

4.1.2 Implementation details

On all the dataset, we set the dimension of the transformed features to 200 and train our DAML model using three fully connected layers for both image and text modalities. 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.

During the training procedure, the batch size is set to 64 for our DAML on all datasets. We tune the model parameters α, β, σ using grid search (for each parameter in range of [0.001, 100] with 10 times per step). In our experiment, the three parameters are empirically set to be 0.01, 0.1 and 1.0, respectively, which show stable performance on different datasets. In addition, to make fair evaluation with the state-of-the-art methods, we not only

[3] http://vision.cs.uiuc.edu/pascal-sentences/

refer to the published results in the corresponding papers but also re-evaluate some of those methods implementations provided by respective authors to obtain objective assessment.

4.1.3 Evaluation metric

We apply the proposed DAML to two cross-modal 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, 5, 22, 35]. To calculate mAP, we first evaluate the average precision (AP) of the retrieval result for each query then average the AP values over the query set. We implement the proposed model using Tensorflow and run the experiments on a desktop machine with 4-core CPU at 4 GHz, 32 GB memory and Geforece Titan X GPU.

4.2 Comparison with existing methods

We first compare our DAML approach with 10 state-of-the-art methods on Wikipedia dataset, which has been widely adopted as a benchmark dataset in the literature. The compared methods are: 1) CCA [10], CCA-3V [7], LCFS [35], JRL [47] and JFSSL [34], which are traditional cross-modal retrieval methods; and 2) Multimodal-DBN [30], Bimodal-AE [18], Corr-AE [5], and CMDN [19], which are DNN based.

Table 2 shows the mAP of our DAML and the compared methods on the Wikipedia dataset using shallow and deep features, respectively. From Table 2, we can draw the follow observations: 1) Our DAML significantly outperforms both the traditional and the DNN based cross-modal retrieval methods. Especially, comparing to CMDN which gets the best retrieval accuracy in all the compared methods, our DAML further gains improvement by 4.66% and 5.05% in average using shallow and deep features, respectively. It is worth mention that CMDN also model inter-modal invariance and intra-modal discrimination jointly in multi-task learning framework, while the adversarial learning facilitates our DAML well balance inter-modal invariance and intra-modal discrimination to obtain more

Table 2 Cross-modal retrieval comparison on Wikipedia dataset. Here "–" denotes that no experimental results with same settings are available

Methods	Shallow feature			Deep feature		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [22]	0.255	0.185	0.220	0.267	0.222	0.245
Multimodal DBN [30]	0.149	0.150	0.150	0.204	0.183	0.194
Bimodal-AE [18]	0.236	0.208	0.222	0.314	0.290	0.302
CCA-3V [7]	0.275	0.224	0.249	0.437	0.383	0.410
LCFS [35]	0.279	0.214	0.246	0.455	0.398	0.427
Corr-AE [5]	0.280	0.242	0.261	0.402	0.395	0.398
JRL [47]	0.344	0.277	0.311	0.453	0.400	0.426
JFSSL [34]	0.306	0.228	0.267	0.428	0.396	0.412
CMDN [19]	–	–	–	0.488	0.427	0.458
DCML [15]	0.352	0.261	0.307	0.526	0.463	0.495
DAML (Proposed)	0.356	0.267	0.322	0.559	0.481	0.520

effective cross-modal representation. 2) Our DAML is superior to CCA, Bimodal-AE, Corr-AE, CMDL and CMDN that use the correlation loss based on coupled samples to model the inter-modal similarity. The reason is that the proposed double triplet constraints are effective to leverage the cues of both similar and dissimilar pairs relying on their discriminant labels, which benefits DAML to effectively model the inter-modal similarity. It consistently indicates that our DAML is more effective to explore to inter-modal similarity than DCML. 3) Our DAML is also outperforms LCFS, CDLFM, LGCFL, JRL, JFSSL that also leverage class label information to model the intra-modal discrimination. Different from these methods, our DAML formulates the feature discrimination and correlation loss that model the inter-modal invariance and intra-modal discrimination, which jointly obtain better category separation across different modalities.

Figure 2 shows three examples of text queries and the top five images retrieved by the proposed DAML for the *Text2Img* task on Wiki dataset. It can be observed that our method finds the closet matches of the image modality at the semantic level for both text queries. And the retrieved images are all belonging to the same label of the text queries, i.e., "warfare" and "literature" respectively.

Moreover, the retrieval results on Pascal Sentence dataset and NUS-WIDE-10k dataset are shown in Table 3. We can see that the our DAML consistently achieves the best performance compared to its counterparts. Specifically, our DAML outperforms the best counterpart CMDN in terms of mAP score by 0.001 and 0.017 on average.

4.3 Further analysis on DAML

4.3.1 Visualization of learned adversarial representation

We further investigate the effectiveness of the cross-modal representations learned by our DAML. 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.

Figure 2 Typical examples of the *Text2Img* task obtained by our proposed DAML on Wiki dataset with CNN features. In each example, the text query and the top five images retrieved are listed in the following columns

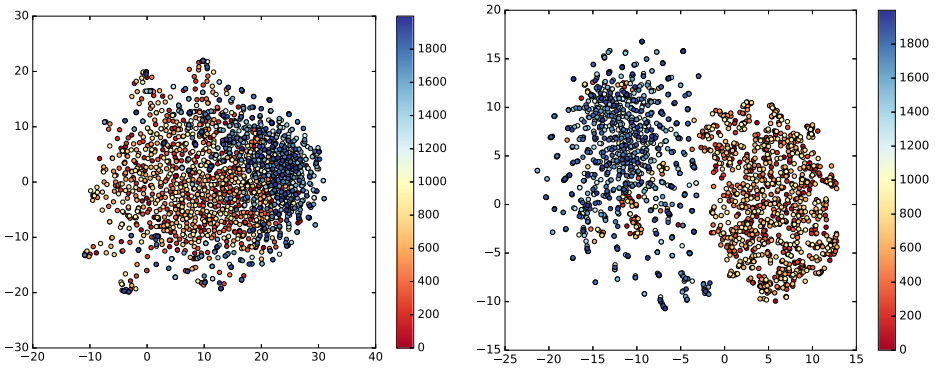
Table 3 Cross-modal retrieval comparison in terms of mAP on Pascal Sentences and NUSWIDE-10k dataset. Here "–" denotes that no experimental results with same settings are available

Methods	Pascal Sentences			NUSWIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [22]	0.363	0.219	0.291	0.189	0.188	0.189
Multimodal DBN [30]	0.477	0.424	0.451	0.201	0.259	0.230
Bimodal-AE [18]	0.456	0.470	0.458	0.327	0.369	0.348
LCFS [35]	0.442	0.357	0.400	0.383	0.346	0.365
Corr-AE [5]	0.489	0.444	0.467	0.366	0.417	0.392
JRL [47]	0.504	0.489	0.496	0.426	0.376	0.401
CMDN [19]	0.534	0.534	0.534	0.492	0.515	0.504
DCML [15]	–	–	–	0.514	0.468	0.491
DAML (Proposed)	0.531	0.539	0.535	0.512	0.534	0.523

Figure 3 shows the t-SNE embedding for the data distribution of Wiki dataset. Figure 3a shows the features with adversarial loss and Figure 3b shows the same without adversarial loss. We can see that without adversarial loss, the transformed features in Figure 3a are still scattered and the adversarial loss indeed effectively closes the gap between different modalities. In Figure 3b, the transformed features are likely to form a single cluster. 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.

4.3.2 Balance of label predicting and structure preserving

Furthermore, the adversarial learning in our DAML is also beneficial to balance the processes of feature discrimination and feature correlation, which model intra-modal discrimination and inter-modal invariance, respectively. To investigate the contributions of these two processes, we develop two variations of DAML: DAML with feature discrimination loss L_{dis} only, and DAML with feature correlation $\mathcal{L}_{cor}$ only. The optimization



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

Figure 3 t-SNE visualization for the chosen data in Wiki. Red represents visual features and blue represents text features

Table 4 Performance of cross-modal retrieval with full DAML method, DAML method with $\mathcal{L}_{dis}$ only, and DAML method with $\mathcal{L}_{cor}$ only

Methods	Wikipedia			Pascal sentences		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
DAML (with $\mathcal{L}_{dis}$ only)	0.326	0.415	0.371	0.281	0.265	0.273
DAML (with $\mathcal{L}_{cor}$ only)	0.411	0.402	0.407	0.525	0.447	0.486
Full DAML	0.493	0.419	0.456	0.529	0.463	0.496

procedure is similar to DAML. Table 4 shows the performance of DAML and its two variations on Wikipedia dataset and Pascal Sentence dataset. We see that both the intra-modal discrimination and inter-modal invariance terms contribute to the final retrieval rate, indicating that optimizing the L_{dis} term and the $\mathcal{L}_{cor}$ simultaneously performs better than optimizing only one of them. We also see that the intra-modal discrimination term contributes more to the overall performance than the inter-modal invariance term, since in practice the consistent relation across different modalities is difficult to explore.

5 Conclusion

In this paper, we proposed a novel approach Deep Adversarial Metric Learning (DAML) for cross-modal retrieval, which aims to learn discriminative (intra-modality) and invariant (inter-modality) representations in common subspace. We decompose the whole problem into three loss terms: 1) adversarial loss was utilized to minimize the "modality gap" between two unknown distributions of representations from different modalities to promote modality-invariant; 2) for feature discrimination loss, intra-modality similarity was modelled by category information, which ensures learned representations to be discriminative; 3) regarding inter-modality similarity, we utilized feature correlation loss to minimize the distances among intra-class cross-modality samples and maximize the distances among inter-class cross-modality samples. The experimental results on three widely used multimodal datasets show the proposed DAML outperforms several state-of-art methods on cross-modal retrieval tasks.

Acknowledgements This work is partially supported by NSFC grant No. 61602089, No. 61673088, No. 61502080; the 111 Project No. B17008; the Fundamental Research Funds for Central Universities ZYGX2016KYQD114; the LEADER of MEXT-Japan (16809746); the Telecommunications Foundation; the REDAS and SCAT.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML, pp. 1247–1255 (2013)
2. Chu, L., Zhang, Y., Li, G., Wang, S., Zhang, W., Huang, Q.: Effective multimodality fusion framework for cross-media topic detection. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 556–569 (2016)
3. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.-T.: Nus-wide: A real-world Web image database from national university of singapore. In: CIVR (2009)
4. Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI **36**(3), 521–535 (2014)

5. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM MM, pp. 7–16. ACM (2014)
6. Ganin, Y., Lempitsky, V.S.: Unsupervised domain adaptation by backpropagation. In: ICML, pp. 1180–1189 (2015)
7. 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)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
9. Gu, Q., Li, Z., Han, J.: Joint feature selection and subspace learning. In: IJCAI (2011)
10. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
11. He, L., Xu, X., Lu, H., Yang, Y., Shen, H.T.: Unsupervised cross modal retrieval through adversarial learning. In: ICME, pp. 1–6 (2017)
12. Kang, C., Liao, S., He, Y., Wang, J., Niu, W., Xiang, S., Pan, C.: Cross-modal similarity learning: a low rank bilinear formulation. In: CKIM, pp. 1251–1260 (2015)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
14. Larsen, A.B.L., Sønderby, S.K., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv:1512.09300 (2015)
15. Liong, V.E., Lu, J., Tan, Y., Zhou, J.: Deep coupled metric learning for cross-modal matching. IEEE Trans. Multimedia **19**(6), 1234–1244 (2017)
16. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.: Adversarial Autoencoders. arXiv, pp. 1–10 (2015)
17. Mignon, A., Jurie, F.: CMML: a new metric learning approach for cross modal matching. In: ACCV (2012)
18. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal deep learning. In: ICML, pp. 689–696 (2011)
19. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: IJCAI, pp. 3846C3853 (2016)
20. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015)
21. Rashtchian, C., Young, M., Hodosh, P., 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 (2010)
22. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM MM, pp. 251–260. ACM (2010)
23. Sharma, A., Kumar, A., Daume, H., Jacobs, D.: Generalized multiview analysis: a discriminative latent space. In: CVPR, pp. 2160–2167. IEEE (2012)
24. Siena, S., Boddeti, V.N., Kumar, B.V.K.V.: Coupled marginal fisher analysis for low-resolution face recognition. In: ECCV workshops and demonstrations, pp. 240–249 (2012)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, arXiv:1409.1556 (2014)
26. Song, J., Gao, L., Liu, L., Zhu, X., Sebe, N.: Quantization-based hashing: a general framework for scalable image and video retrieval. Pattern Recogn. **75**, 175–187 (2018)
27. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: ACM SIGMOD, pp. 785–796 (2013)
28. Song, J., Zhang, H., Li, X., Gao, L., Wang, M., Hong, R.: Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Trans. Image Process. **25**(11), 4999–5011 (2018)
29. Song, Y., Wang, W., Zhang, A.: Automatic annotation and retrieval of images. World Wide Web **6**(2), 209–231 (2003)
30. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: ICML workshop (2012)
31. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: NIPS, pp. 2222–2230 (2012)
32. Tian, Q., Chen, S.: Cross-heterogeneous-database age estimation through correlation representation learning. Neurocomput. **238**(C), 286–295 (2017)
33. Wang, J., Zhang, T., Song, J., Sebe, N., Shen, H.T.: A survey on learning to hash. IEEE Trans. Pattern Anal. Mach. Intell. **26**(99), 15–29 (2017)
34. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. TPAMI **38**(10), 2010–2023 (2011)
35. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV, pp. 2088–2095 (2013)

36. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A Comprehensive Survey on Cross-modal Retrieval. arXiv, pp. 1–20 (2016)
37. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. VLDB J **25**(1), 79–101 (2016)
38. Wang, X., Gao, L., Wang, P., Sun, X., Liu, X.: Two-stream 3d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Trans. Multimedia **PP**(99), 1–1 (2017)
39. Wu, Y., Wang, S., Huang, Q.: Online asymmetric similarity learning for cross-modal retrieval. In: CVPR, pp. 3984–3993 (2017)
40. Xu, X., He, L., Shimada, A., Taniguchi, R., Lu, H.: Learning unified binary codes for cross-modal retrieval via latent semantic hashing. Neurocomputing **213**, 191–203 (2016)
41. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans. Image Process. **26**(5), 2494–2507 (2017)
42. Xu, X., Shimada, A., Taniguchi, R., He, L.: Coupled dictionary learning and feature mapping for cross-modal retrieval. In: ICME, pp. 1–6. IEEE (2015)
43. Xu, X., Yang, Y., Shimada, A., Taniguchi, R., He, L.: Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In: ACM MM, pp. 847–850. ACM (2015)
44. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: CVPR, pp. 3441–3450 (2015)
45. Yang, Y., Zha, Z.-J., Gao, Y., Zhu, X., Chua, T.-S.: Exploiting Web images for semantic video indexing via robust sample-specific loss. IEEE Trans. Multimedia **16**(6), 1677–1689 (2014)
46. Yang, Y., Zhang, H., Zhang, M., Shen, F., Li, X.: Visual coding in a semantic hierarchy. In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 59–68, ACM (2015)
47. Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans. Circuits Syst. Video Technol. **24**, 965C–978 (2014)
48. Zhang, H., Gao, X., Wu, P., Xu, X.: A cross-media distance metric learning framework based on multi-view correlation mining and matching. World Wide Web **19**(2), 181–197 (2016)
49. Zhuang, Y., Wang, Y., Wu, F., Zhang, Y., Lu, W.: Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: AAAI (2013)
50. Zou, F., Chen, Y., Song, J., Zhou, K., Yang, Y., Sebe, N.: Compact image fingerprint via multiple kernel hashing. IEEE Transaction on Multimedia **17**(7), 1006–1018 (2015)