Adversarial Cross-Modal Retrieval

Bokun Wang
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
csbkwang@gmail.com

Yang Yang*
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
dlyyang@gmail.com

Xing Xu
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
xing.xu@uestc.edu.cn

Alan Hanjalic
Multimedia Computing Group, Delft University of Technology
a.hanjalic@tudelft.nl

Heng Tao Shen
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China
shenhengtao@hotmail.com

ABSTRACT

Cross-modal retrieval aims to enable flexible retrieval experience across different modalities (*e.g.*, texts *vs.* images). The core of cross-modal retrieval research is to learn a common subspace where the items of different modalities can be directly compared to each other. In this paper, we present a novel Adversarial Cross-Modal Retrieval (ACMR) method, which seeks an effective common subspace based on adversarial learning. Adversarial learning is implemented as an interplay between two processes. The first process, a feature projector, tries to generate a modality-invariant representation in the common subspace and to confuse the other process, modality classifier, which tries to discriminate between different modalities based on the generated representation. We further impose triplet constraints on the feature projector in order to minimize the gap among the representations of all items from different modalities with same semantic labels, while maximizing the distances among semantically different images and texts. Through the joint exploitation of the above, the underlying cross-modal semantic structure of multimedia data is better preserved when this data is projected into the common subspace. Comprehensive experimental results on four widely used benchmark datasets show that the proposed ACMR method is superior in learning effective subspace representation and that it significantly outperforms the state-of-the-art cross-modal retrieval methods.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**;

KEYWORDS

cross-modal retrieval; adversarial learning; modality gap

*Corresponding author.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123326

1 INTRODUCTION

In order to maximally benefit from the abundance 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 one if they are semantically related, regardless of the type of modalities (*e.g.* text, visual or audio) of the items. Modeling of similarity links has traditionally focused mainly on single-modality scenarios. For instance, information retrieval has focused on bringing similar textual documents together, while content-based image, video or audio retrieval has attempted the same for images, videos and audio items, respectively. In order to provide an answer to the above challenge, research towards reliable solutions for cross-modal applications [37, 38], *e.g. cross-modal retrieval* [32], which operate across modality boundaries, has gained significant momentum recently.

Since features of different modalities usually have inconsistent distribution and representation, a *modality gap* needs to be bridged, that is, ways need to be found to assess semantic similarity of items across modalities. A common approach to bridge the modality gap is *representation learning*. The goal there is to find (*i.e.*, learn) projections of data items from different modalities into a common (modality agnostic) feature representation subspace in which the similarity between them can be assessed directly. A variety of cross-modal retrieval methods [5, 9, 21, 30, 31] have been proposed recently, which propose different ways of learning the common representation subspace. For example, early works, like CCA-based methods [15, 39] and graph-based methods [40, 43], learn linear projections to generate a common representation by maximizing the cross-modal pairwise item correlation or item classification accuracy.

With the rapid development of deep neural network (DNN) models that have provided scalable nonlinear transformations for effective feature representations in single-modal scenarios, such as image classification, DNN has increasingly been deployed in the cross-modal retrieval context as well, and then in particular to exploit nonlinear correlations when learning a common subspace

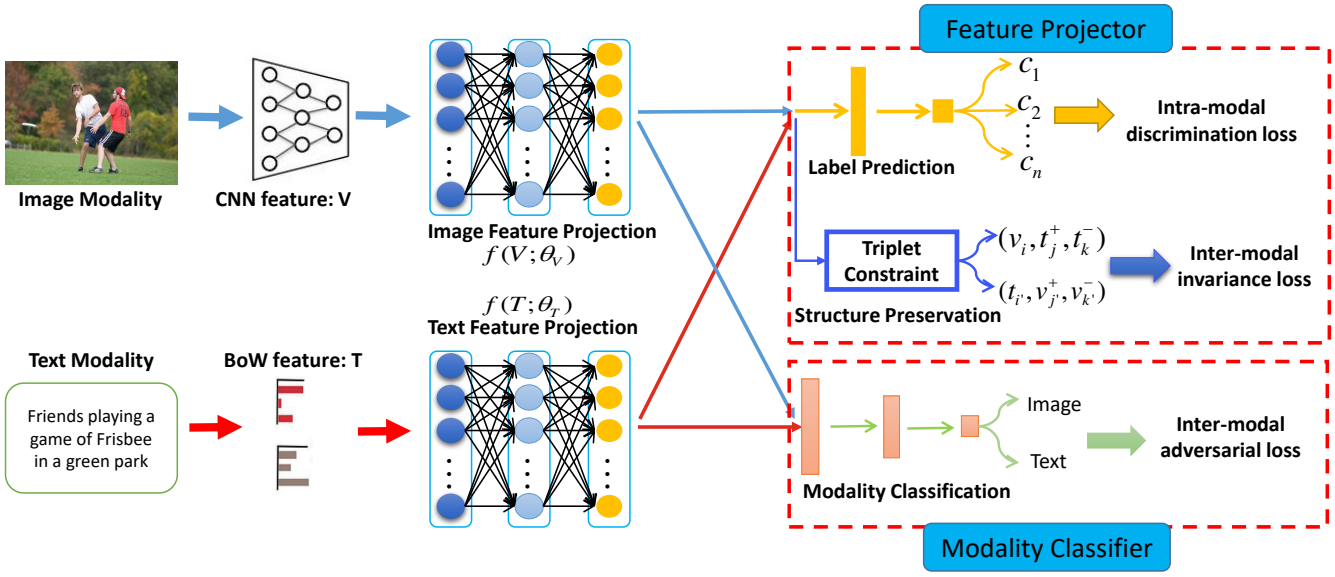


Figure 1: The general flowchart of the proposed ACMR method. It is built around the minimax game involving two processes (alorithmic modules) as "players": a modality classifier distinguishing the items in terms of their modalities, and a feature projector generating modality-invariant and discriminative representations and aiming to confuse the modality classifier.

[1, 5, 20, 21, 36]. The existing DNN-based cross-media retrieval models typically focus solely on preserving the pairwise similarity of the coupled cross-modal items (*e.g.*, an image and a piece of text) that share semantic labels and serve as input in the model learning process. However, for one item of one modality, there may exist more than one semantically different items of the same modality so that this focus on pairwise coupled items only is far from sufficient. Therefore, a common representation learned in this way fails to fully preserve the underlying cross-modal semantic structure in data. Preserving this structure would require that the gap among the representations of all items from different modalities with same semantic labels is minimized (*e.g.*, linking any text and any image on the same topic together), while the distances among semantically different items of the same modality are maximized (*e.g.*, separating two images or two texts from each other if they are not related).

We propose to address this drawback of the existing DNN-based cross-media retrieval methods by a new framework that we refer to as *Adversarial Cross-Model Retrieval* (ACMR) and that is built around the concept of *adversarial learning* [8]. As illustrated in Figure 1, the core of the framework is the interplay between two processes, a *feature projector* and a *modality classifier*, conducted as a minimax game. A feature projector performs the main task of representation learning, namely, generating a modality-invariant representation for items from different modalities in the common subspace. It has the objective to confuse a modality classifier that acts as an adversary. Modality classifier tries to distinguish the items in terms of their modalities and in this way steers the learning of the feature projector. By bringing the modality classifier in the adversary role, it is expected that the modality invariance is reached more efficiently, but also more effectively, through reaching a better alignment of distributions of item representations across modalities. The representation subspace being optimal for cross-modal retrieval will then result through the convergence of this process, namely when the modality classifier "fails". Furthermore, the feature projector is learned such that it jointly performs label prediction and preserves the underlying cross-modal semantic structure in data. In this way it can ensure that the learned representation is both discriminative within a modality and invariant across modalities. The latter is achieved by imposing more constraints on inter-modal item relations than in previously proposed methods, which only focus on pairwise item correlation [1, 5, 21].

The proposed ACMR method was evaluated on four (three small-scale and one relatively large-scale) benchmark datasets and using many existing methods as references. The experimental results show that it significantly outperforms the state-of-the-art in cross-modal retrieval. In Section 2, we position our approach in the context of the related existing work. Then, in Section 3, we describe our ACMR method in detail and evaluate it experimentally in Section 4. Section 5 concludes the paper.

2 RELATED WORK

The main contribution of this paper concerns the representation learning component of a cross-modal retrieval framework. Representation learning has been approached in different ways, depending on the type of information that is used for learning, the type of the targeted representation and the learning approaches being deployed. Two general categories of the representation learning approaches can be distinguished, *real-valued* and *binary* representation learning. The binary approaches, also referred to as *cross-modal hashing* are more geared towards retrieval efficiency and aim at mapping the items of different modalities into a common binary Hamming space [24, 27, 34, 42]. Since the focus is on efficiency,

concessions typically need to be made regarding retrieval accuracy (effectiveness).

The approach proposed in this paper falls in the category of real-valued approaches. In this category, several subclasses of approaches can be distinguished: unsupervised [1, 5, 9, 29, 36], pairwise [5, 11, 25, 43], ranking-based [33, 39] and supervised [7, 30, 31, 40] ones. With ACMR, we combine for the first time the concepts of supervised representation learning for cross-modal retrieval and adversarial learning. Our approach is motivated, on the one hand, by the deficiencies we see in most of the (un)supervised methods, and in particular regarding the effectiveness of the learning process (focusing on individual pairs of samples) and learning objectives (typically the variants of the *correlation loss* [5, 7]) they deploy. On the other hand, our approach is inspired by the ideas proposed in some ranking-based approaches, especially regarding the deployment of a *triplet ranking loss* [33, 39] as learning objective, which has been found effective for achieving the main goal of representation learning, namely intra-modal discriminativeness and inter-modal invariance. Furthermore, our approach was inspired by the effectiveness of adversarial learning for various applications, like learning discriminative image features [17], or (un)supervised domain adaptation to enforce domain-invariant features [2, 6, 41], and regularizing correlation loss between cross-modal items [10].

3 PROPOSED METHOD

3.1 Problem Formulation

Without losing generality, we focus on cross-modal representation learning for bimodal data, specifically for images and text. We assume that there is a collection of n instances of image-text pairs, denoted as $O = \{o_i\}_{i=1}^{n}$, $o_i = (v_i, t_i)$, where $v_i \in \mathbb{R}^{d_v}$ is an image feature vector and $\mathbf{t}_i \in \mathbb{R}^{d_t}$ is a text feature vector. d_v and d_t are the feature dimensions with, usually, $d_v \neq d_t$. In addition, each instance o_i is also assigned a semantic label vector $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{ic}] \in \mathbb{R}^c$, where c is the total number of semantic categories. If the ith instance belongs to the jth semantic category, $y_{ij} = 1$, otherwise $y_{ij} = 0$. Note that o_i can belong to a single, but also multiple semantic categories. We denote the image feature matrix, text feature matrix and label matrix for all instances in O as $\mathcal{V} = \{v_1, ..., v_n\} \in \mathbb{R}^{d_v \times n}$, $\mathcal{T} = \{t_1, ..., t_n\} \in \mathbb{R}^{d_t \times n}$ and $\mathcal{Y} = \{y_1, ..., y_n\} \in \mathbb{R}^{c \times n}$, respectively.

As the image features $\mathcal{V}$ and text features $\mathcal{T}$ typically have different statistical properties and follow unknown (complex) distributions, they cannot be directly compared against each other for cross-modal retrieval. To make an image and a text directly comparable, we aim at finding a common subspace $\mathcal{S}$ the image features $\mathcal{V}$ and text features $\mathcal{T}$ can be projected to as $\mathcal{S}_\mathcal{V} = f_\mathcal{V}(\mathcal{V}; \theta_\mathcal{V})$ and $\mathcal{S}_\mathcal{T} = f_\mathcal{T}(\mathcal{T}; \theta_\mathcal{T})$. Here, $f_\mathcal{V}(v; \theta_\mathcal{V})$ and $f_\mathcal{T}(t; \theta_\mathcal{T})$ are the mapping functions and $\mathcal{S}_\mathcal{V} \in \mathbb{R}^{m \times n}$ and $\mathcal{S}_\mathcal{T} \in \mathbb{R}^{m \times n}$ are the transformed features of an image and a text in $\mathcal{S}$, respectively.

Specific to the ACMR method proposed in this paper is that we aim at learning more effective transformed features $\mathcal{S}_\mathcal{V}$ and $\mathcal{S}_\mathcal{T}$ in $\mathcal{S}$ for different modalities. As argued earlier, we require from the distributions of $\mathcal{S}_\mathcal{V}$ and $\mathcal{S}_\mathcal{T}$ to be modality-invariant and semantically discriminative, but also to better preserve the underlying cross-modal similarity structure in data. We explain in the following subsections how these requirements are met.

3.2 Adversarial Cross-Modal Retrieval

The general framework of the proposed ACMR method is shown in Figure 1. For simplicity, we assume that the features $\mathcal{V}$ and $\mathcal{T}$ have already been extracted from images and text, respectively. Image and text features first pass through respective transforms $f_\mathcal{V}$ and $f_\mathcal{T}$, which are conceptually inspired by the existing subspace learning methods [9, 31, 36] and in our case realized as feed-forward networks. The fully-connected layers have abundant parameters to ensure enough capacity of representations considering large margin of statistical properties between the image and text modality. Then, in the second step, the minimax game "played" between two processes, the feature projector and modality classifier, is introduced to steer the representation learning. We model these processes and their interaction such to effectively and efficiently meet the requirements defined above.

3.3 Modality Classifier

We first define a modality classifier D with parameters θ_D, which acts as "discriminator" in GAN [8]. The projected features from an image are assigned the label $\overline{01}$, while the projected features from a text are assigned the label $\overline{10}$. For the modality classifier, the goal is to detect the modality of an item as reliably as possible given an unknown feature projection. For the classifier implementation, we used a 3-layer feed-forward neural network with parameters θ_D (see section 4.1 for implementation details).

In the ACDM method, the modality classifier acts as an adversary. Therefore, we refer to the classification loss this process tries to minimize as *adversarial loss*. The adversarial loss L_{adv} can now formally be defined as:

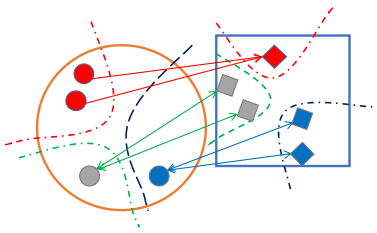
$$\mathcal{L}_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{m_i} \cdot (\log D(v_i; \theta_D) + \log(\mathbf{1} - D(t_i; \theta_D)))). \quad (1)$$

Essentially, $\mathcal{L}_{adv}$ denotes the cross-entropy loss of modality classification of 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.

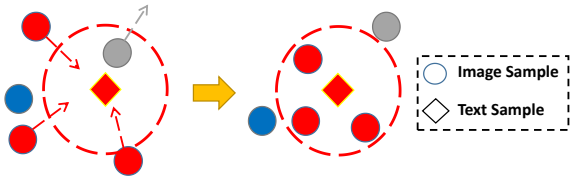
3.4 Feature Projector

The correlation loss, like in most of the existing work, targets solely that the correlation of the individual pairs of semantically coupled cross-modal items is preserved in the new representation subspace. As already discussed earlier in the paper, this is not sufficient because semantic matching may involve more than two items. Besides, correlation loss also fails to differentiate between semantically different items of the same modality. This leads to feature representations that are not discriminative enough and will limit the performance of cross-modal retrieval.

In view of the above, we propose to model the feature projector, which embodies the process of modality-invariant embedding of texts and images into a common subspace, as a combination of two steps: label prediction and structure preservation. The label prediction process enables the projected feature representations for each modality in the common subspace to be discriminative given the semantic labels. The structure preservation process ensures that the feature representations belonging to the same semantic



(a) Illustration of the joint effect of label prediction and structure preservation.



(b) General idea of our proposed ACMR to achieve an improved, modality-invariant subspace embedding with cross-modal semantic structure preservation.

Figure 2: Illustration of the basic idea underlying the proposed ACMR method. Images texts are represented by squares and circles, respectively. Semantically related cross-modal items are indicated by the same color.

label is invariant across modalities. The joint effect of these two processes is illustrated in Figure. 2(a). There, each circle represents an image and each rectangle a textual item. Furthermore, the circles and rectangles of the same color belong to the same semantic category. The process leading to this effect is illustrated in Figure 2(b). In the remainder of this section, we describe in detail the label prediction and structure preservation modules, underlying the subspace embedding process.

3.4.1 Label Prediction. 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. We use these probability distributions $\hat{p}$ to formulate the intra-modal discrimination loss as follows:

$$\mathcal{L}_{imd}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot (\log \hat{p}_i(v_i) + \log \hat{p}_i(t_i))). \quad (2)$$

Similar to the inter-modal adversarial loss in Eq. 1, $\mathcal{L}_{imd}$ denotes the cross-entropy loss of semantic category classification of all the instances $o_i, i = 1, ..., n$. Here, θ_{imd} denotes the parameters of the classifier, n is the number of instances within each mini-batch, $\mathbf{y}_i$ is the groundtruth of each instance, while $\hat{p}_i$ is the generated probability distribution per item (image or text) of the instance o_i.

3.4.2 Structure Preservation. In order to ensure the preservation of inter-modal invariance, we aim at minimizing the gap among the representations of all semantically similar items from different modalities, while maximizing the distance between semantically different items of the same modality. Inspired by the ranking-based cross-media retrieval approaches [33, 39], we enforce *triplet constraints* onto the embedding process via a triplet loss term we formulated for this purpose.

Instead of applying a computationally expensive scheme that samples triplets in the whole instance space, we performed triplet sampling from labeled instances in each mini-batch. Firstly, all samples from different modalities, but with the same label, were built as coupled samples from the perspective of both images and text samples. In other words, we built couples of the form $\{(v_i, t_i^+)\}_i$, where image is selected as anchor while the text with same label is assigned as a positive match, and also the couples of the form $\{(t_i, v_i^+)\}_i$ with a text item as an anchor and an image as a positive match.

Secondly, all distances between the mapped representations $f_{\mathcal{V}}(\mathcal{V}; \theta_{\mathcal{V}})$ and $f_{\mathcal{T}}(\mathcal{T}; \theta_{\mathcal{T}})$ per coupled item pair were computed and sorted using the ℓ_2 norm:

$$\ell_2(v, t) = \|f_{\mathcal{V}}(v; \theta_{\mathcal{V}}) - f_{\mathcal{T}}(t; \theta_{\mathcal{T}})\|_2 . \quad (3)$$

Then, we also select negative samples from unmatched image-text pairs having different semantic labels to build the sets of triplet samples per semantic label l_i: $\{(v_i, t_i^+, t_j^-)\}_i$ and $\{(t_i, v_i^+, v_j^-)\}_i$. In this way of sampling, we can ensure that non-empty triplet sample sets will be constructed independently of how samples in the original dataset were organized into the mini-batches.

Finally, we compute the inter-modal invariance loss across image and text modalities using the following expressions that take as input the sample sets $\{(v_i, t_j^+, t_k^-)\}_i$ and $\{(t_i, v_j^+, v_k^-)\}_i$, respectively:

$$\mathcal{L}_{imi,\mathcal{V}}(\theta_{\mathcal{V}}) = \sum_{i,j,k}(\ell_2(v_i, t_j^+) + \lambda \cdot \max(0, \mu - \ell_2(v_i, t_k^-)), \quad (4)$$

$$\mathcal{L}_{imi,\mathcal{T}}(\theta_{\mathcal{T}}) = \sum_{i,j,k}(\ell_2(t_i, v_j^+) + \lambda \cdot \max(0, \mu - \ell_2(t_i, v_k^-)). \quad (5)$$

Then the overall inter-modal invariance loss can now be modeled as a combination of $\mathcal{L}_{imi,\mathcal{V}}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}})$ and $\mathcal{L}_{imi,\mathcal{T}}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}})$:

$$\mathcal{L}_{imi}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}}) = \mathcal{L}_{imi,\mathcal{V}}(\theta_{\mathcal{V}}) + \mathcal{L}_{imi,\mathcal{T}}(\theta_{\mathcal{T}}). \quad (6)$$

In addition, the regularization term below is introduced to prevent the learned parameters from overfitting, where F denotes the Frobenius norm and W_v^l, W_t^l represent the layer-wise parameters of DNNs.

$$\mathcal{L}_{reg} = \sum_{l=1}^{L}(||W_v^l||_F + ||W_t^l||_F). \quad (7)$$

3.4.3 Feature projector. Based on the above, the loss function of the feature projector, referred to as *embedding loss*, is formulated as the combination of the intra-modal discrimination loss and the inter-modal invariance loss with regularization:

$$\mathcal{L}_{emb}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}) = \alpha \cdot \mathcal{L}_{imi} + \beta \cdot \mathcal{L}_{imd} + \mathcal{L}_{reg}, \quad (8)$$

where the hyper-parameters α and β control the contributions of the two terms.

3.5 Adversarial Learning: Optimization

The process of learning the optimal feature representation is conducted by jointly minimizing the adversarial and embedding losses, as obtained in Eq. 1 and Eq. 8, respectively. Since the optimization goals of these two objective functions are opposite, the process runs as a **minimax game** [8] of the two concurrent sub-processes:

$$(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) = \underset{\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}}{\arg\min} \left(\mathcal{L}_{emb}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}) - \mathcal{L}_{adv}(\hat{\theta}_D)\right), \quad (9)$$

$$\hat{\theta}_D = \underset{\theta_D}{\arg\max} \left(\mathcal{L}_{emb}(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) - \mathcal{L}_{adv}(\theta_D)\right). \quad (10)$$

This minimax game can be implemented using a stochastic gradient descent optimization algorithms, like Adam [13]. As proposed in [6], minimax optimization can be performed efficiently by incorporating Gradient Reversal Layer (GRL), which is transparent when forward-propagating, but which multiplies its value by -1 when back-propagating. If Gradient Reversal layer is added before the first layer of the modality classifier, the minimax optimization can be performed simultaneously, as shown in the Algorithm 1.

Algorithm 1 Pseudocode of optimizing our ACMR.

Initialization: Image features for current batch: $\mathcal{V} = \{v_1, ..., v_n\}$;
Text features for current batch, $\mathcal{T} = \{t_1, ..., t_n\}$;
Corresponding labels for current batch, $\mathcal{Y} = \{y_1, ..., y_n\}$;
hyperparameters: k, λ, α, β;
m samples in minibatch for each modality;

update until convergence:

1: **for** k steps **do**
2: update parameters $\theta_{\mathcal{V}}$, $\theta_{\mathcal{T}}$ and θ_{imd} by **descending** their stochastic gradients:
3: $\theta_{\mathcal{V}} \leftarrow \theta_{\mathcal{V}} - \mu \cdot \nabla_{\theta_{\mathcal{V}}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
4: $\theta_{\mathcal{T}} \leftarrow \theta_{\mathcal{T}} - \mu \cdot \nabla_{\theta_{\mathcal{T}}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
5: $\theta_{imd} \leftarrow \theta_{imd} - \mu \cdot \nabla_{\theta_{imd}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
6: **end for**
7: update parameters of modality classifier by **ascending** its stochastic gradeints through Gradient Reversal Layer:
8: $\theta_D \leftarrow \theta_D + \mu \cdot \lambda \cdot \nabla_{\theta_D} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
9: **return** learned representations in common subspace: $f_{\mathcal{V}}(\mathcal{V})$ and $f_{\mathcal{T}}(\mathcal{T})$

4 EXPERIMENTS

We conduct experiments on four widely-used cross-modal datasets: Wikipedia dataset [4], NUS-WIDE-10k dataset [3], Pascal Sentence dataset [23], and MSCOCO dataset [16]. For the first three datasets, each image-text pair is linked by a single class label and the text modality consists of discrete tags. In the last dataset, MSCOCO, each image-text pair is associated to multiple class labels and the text modality consists of sentences. In the experiments reported below, we first compare our proposed ACMR method with the state-of-the-art methods to verify its effectiveness. Then we conduct additional evaluations in order to investigate the performance of ACMR in more detail.

4.1 Experimental Setup

4.1.1 Datasets and Features. Here we briefly introduce the four datasets mentioned above. For a fair comparison, we exactly follow the dataset partition and feature extraction strategies from [22, 32]. The statistics of the four datasets are summarized in Table 1.

Table 1: General statistics of the four datasets used in our experiments, where "*/*" in the "Instances" column 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
MSCOCO	66,226/16,557	500	4,096d VGG	3,000d BoW

Since the image feature extracted from a deep Convolutional Neural Network (CNN) has been widely used for image representation, we also adopt this deep feature to represent images in all datasets for our experiments. Specifically, the adopted deep feature is a 4,096d vector extracted by the fc7 layer of VGGNet [26]. For representing text instances, we use a well-known bag-of-words (BoW) vector with the TF-IDF weighting scheme and with the dimension in each dataset indicated in Table 1. In addition, in order to enable a fair comparison to several earlier cross-modal retrieval approaches evaluated on the Wikipedia dataset, we also adopt the publicly available 128d SIFT feature for images and 10d LDA feature for text representation.

4.1.2 Implementation Details. We deploy three-layer feed-forward neural networks activated by *tanh* function to nonlinearly project the raw image and text features into a common subspace, *i.e.*, ($V \rightarrow 2000 \rightarrow 200$ for image modality and $T \rightarrow 500 \rightarrow 200$ for text modality). For the modality classifier, we stick to the three fully connected layers ($f \rightarrow 50 \rightarrow 2$). Furthermore, Softmax activation was added after the last layer of the semantic classifier (Section 3.3.1) and modality classifier.

Regarding the parameters of the Algorithm 1, the batch size is set to 64 and k is empirically set to be 5. After fixing the value of λ at 0.05, we tuned the model parameters α and β using grid search (in both cases from 0.01 to 100, 10 times per step). The analysis of α and β is displayed in Figure 6(a). The best reported results of ACMR are obtained for the optimal values of α and β per dataset. In addition, for a fair comparison with the state-of-the-art methods, we not only refer to the published results in the corresponding papers but also re-evaluate some of those methods with the provided implementation codes in order to achieve a comprehensive assessment.

4.1.3 Evaluation Metric. The evaluation of the results of all experiments is done in terms of the mean average precision (mAP), which is a classical performance evaluation criterion in the research on cross-modal retrieval [5, 21]. Specifically, and similarly to [5],

Methods	Shallow feature			Deep feature		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA	0.255	0.185	0.220	0.267	0.222	0.245
Multimodal 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
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
JFSSL	0.306	0.228	0.267	0.428	0.396	0.412
CMDN	-	-	-	0.488	0.427	0.458
ACMR (Proposed)	**0.366**	**0.277**	**0.322**	**0.619**	**0.489**	**0.546**

Table 2: Comparison of the cross-modal retrieval performance on the Wikipedia dataset. Here, "-" denotes that no experimental results with same settings are available.

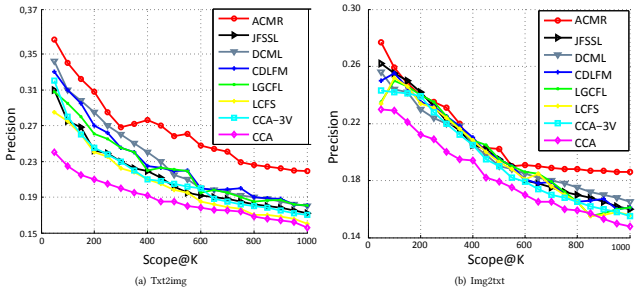


Figure 3: Precision-scope curves on the Wiki dataset for the Img2Txt and Txt2Img experiments with K ranges from 50 to 1000.

we computed the mAP on the ranked lists of the retrieved results for two different tasks: retrieving text samples using an image query (Img2Txt) and retrieving images using a text query (Txt2Img). Besides, we also display the precision-scope curves (like in [31, 35]) for the proposed ACMR method and all reference methods, where the scope is specified by the number of the top-ranked texts/images presented to the users, varying from 1 to 1000.

4.2 Comparison with Existing Methods

We first compare our ACMR approach with 9 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 [9], CCA-3V [7], LCFS [31], JRL [40] and JFSSL [30], which are traditional cross-modal retrieval methods, and 2) Multimodal-DBN [28], Bimodal-AE [20], Corr-AE [5], and CMDN [21], which are DNN based.

Table 2 shows the mAP of our ACMR and the compared methods on the Wikipedia dataset using shallow and deep features. From Table 2, we can draw the follow observations:

(1) Our ACMR significantly outperforms both the traditional and the DNN-based cross-modal retrieval methods. In particular, ACMR outperforms the best competitor, CMDN, by 20.6% and 19.2% in average using shallow and deep features, respectively. While CMDN also models inter-modal invariance and intra-modal discrimination jointly in a multi-task learning framework, this performance improvement clearly shows the advantage of relying on adversarial learning for this purpose.

(2) Our ACMR 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 item similarity. This shows the advantage of using the proposed triplet constraints to leverage the cues of both similar and dissimilar item pairs in learning the representation subspace.

(3) Our ACMR outperforms LCFS, CDLFM, LGCFL, JRL, JFSSL that also leverage class-label information to model the intra-modal discrimination loss. We believe that this is due to the fact that ACMR uses the embedding loss that jointly models the inter-modal invariance and intra-modal discrimination.

The retrieval results on the Pascal Sentence dataset and the NUS-WIDE-10k dataset are shown in Table 3. We can see that ACMR consistently achieves the best performance compared to its counterparts. The performance improvement on the Pascal Sentences dataset of our method is not as convincing as for the NUSWIDE-10k dataset, which is due to its small scale that prevented us to optimally train a well-performing deep model. However, for the NUSWIDE-10k dataset, our ACMR outperforms the counterparts by 10.6% and 4.47% in image and text query retrieval tasks, respectively, and 7.34% on average. The results also indicate the benefit of using triplet constraints in the multi-label case (NUS-WIDE-10k dataset), since the previous methods tested there solely adopted the pairwise similarity to preserve the inter-modal similarity. The improvement on Pascal Sentences dataset of our method is limited because that dataset is small-scale (only 800 img-txt pairs from 20 categories). Although we utilized some strategies to alleviate overfitting problems (like regularization term, dropout, and early stop), It is still insufficient to train an outperforming deep model.

In addition to the evaluation in terms of the mAP score, we also draw precision-scope curves for additional comparison. Figure 3 shows the curves of ACMR and of CCA, LCFS, JRL, Multimodal-DBN, Bimodal-AE, Corr-AE, and CMDN using the shallow image feature. The precision-scope evaluation is consistent with the mAP scores for both the image and text query tasks, where our ACMR outperforms its counterparts significantly.

The MSCOCO dataset has recently been used for image-to-sentence (Img2Txt) and sentence-to-Image (Txt2Img) retrieval. We use it to compare our ACMR method with several recently proposed methods for the above two tasks, including the traditional method CCA [14] and the DNN-based approaches, such as DVSA [12], m-RNN [19], m-CNN [18] and DSPE [33]. For the evaluation on the MSCOCO dataset, we quote the mAP results obtained by the baselines [12, 14, 18, 19] as referred to [33]. We also follow the evaluation protocol according to these reference methods to ensure fair comparison. The retrieval results of ACMR and the reference methods are listed in Table 4. It is interesting to note that the best performing reference method, DSPE, also uses triplet constraints to preserve the inter-modal data structure in the common subspace. This further strengthens our belief that the choice for the triplet

Methods	Pascal Sentences			NUSWIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA	0.363	0.219	0.291	0.189	0.188	0.189
Multimodal 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 (Proposed)	**0.535**	**0.543**	**0.539**	**0.544**	**0.538**	**0.541**

Table 3: Cross-modal retrieval comparison in terms of mAP on Pascal Sentences and NUSWIDE-10k dataset.

constraints to replace the pairwise correlation loss is the right one. It is reasonable to state that the performance increase by ACMR compared to DSPE is again due to the deployed adversarial learning framework, which boosts learning of a more effective subspace representation, but also due to the integration of triplet constraints for inter-modality invariance and the intra-modal discrimination.

Methods	Img2Txt	Txt2Img	Avg.
CCA (FV HGLMM) [14]	0.791	0.765	0.778
CCA (FV GMM+HGLM) [14]	0.809	0.766	0.788
DVSA [12]	0.805	0.748	0.777
m-RNN [19]	0.835	0.770	0.803
m-CNN [18]	0.841	0.828	0.835
DSPE [33]	0.892	0.869	0.881
ACMR (Proposed)	**0.932**	**0.871**	**0.902**

Table 4: Cross-modal retrieval comparison in terms of mAP on MSCOCO dataset.

4.3 Further Analysis on ACMR

4.3.1 Visualization of the Learned Adversarial Representation. In order to investigate the effectiveness of the learned cross-modal representation of our ACMR, we visualized the distribution of the transformed representations from our trained model on the Wikipedia dataset using t-SNE tool (1000 sample points for each modality). A comparison between Figure 4(a) and Figure 4(b) reveals that adversarial learning has the ability to minimize the modality gap and align distributions of different modalities, *i.e.*, the distributions of text and image modality in Figure 4(b) are better mixed together and less distinguishable from each other. Moreover, our dedicated effort to model intra-modal discriminativeness has shown to further boost the performance. As shown in Figure 4(b) and Figure 4(c), the proposed model not only ensures the alignment of distributions from two modalities, but also efficiently separates sample points into several semantically discriminative clusters, keeping the samples from different modalities in each cluster well aligned.

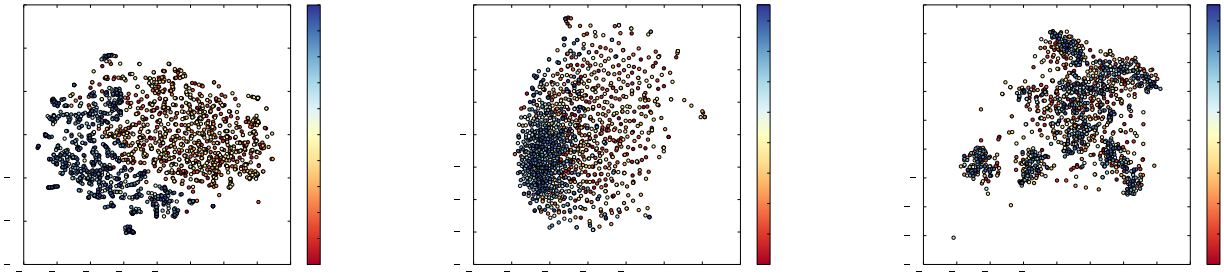
4.3.2 Effect of Adversarial Learning. In our ACMR method, we deploy the adversary principle when jointly optimizing the embedding loss and adversarial loss in the objective function. To further explore the effect of adversarial learning in ACMR, we sampled the values of the embedding loss and the adversarial loss from epoch 1 to 200 and displayed them in Figure 5. The figure shows that during the entire training procedure, the embedding loss decreases almost monotonously and converges smoothly, while the adversarial loss first vibrates (around the initial 10 epochs) and then stabilizes. Notably, the mAP score persistently increases when the adversarial loss is vibrating and holds when the effect of the adversary is fully exploited. The results in Figure 5 are in accordance with the expectation that the modality classifier in our ACMR framework serves as a directional guide for the process of subspace embedding, incorporated in the feature projector. If the value of adversarial loss would explode, modality classifier would fail to direct the process of subspace embedding. Contrary to this, if the adversarial loss were optimized down to zero, modality classifier would win the minimax game, which would mean that the embedding layers fail to generate modality-invariant subspace representations, making cross-modal retrieval impossible.

4.3.3 Effect of combining Label Prediction and Structure Preservation. The feature projection module of our ACMR framework is realized as a combination of two processes, label prediction and structure preservation. In order to investigate in more detail the effect of this combination, we developed and assessed two variations of ACMR: ACMR with L_{imi} only, and ACMR with $\mathcal{L}_{imd}$ only. The optimization procedure in both cases is similar to ACMR. Table 5 shows the performance of ACMR and its two variations on the Wikipedia dataset and the Pascal Sentence dataset. We see that both the intra-modal discriminativeness and inter-modal invariance terms contribute to the final retrieval rate, indicating that optimizing the L_{imi} term and the $\mathcal{L}_{imd}$ simultaneously in our embedding loss model performs better than optimizing only one of them. We also see that the intra-modal discriminativeness 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.

Methods	Wikipedia			Pascal Sentences		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
ACMR (with $\mathcal{L}_{imi}$ only)	0.352	0.430	0.391	0.289	0.274	0.282
ACMR (with $\mathcal{L}_{imd}$ only)	0.425	0.413	0.419	0.533	0.453	0.493
Full ACMR	**0.509**	**0.431**	**0.470**	**0.535**	**0.486**	**0.511**

Table 5: Performance of cross-modal retrieval using the proposed ACMR method, ACMR method with $\mathcal{L}_{imi}$ only, and ACMR method with $\mathcal{L}_{imd}$ only.

4.3.4 Effect of Model Parameters. In previous experiments, we empirically set the model parameters α and β in the objective function of the feature projector (*i.e.*, Eq. 8). As α and β control the contributions of modeling intra-modal discriminativeness and inter-modal invariance, respectively, here we take the Wikipedia dataset with deep features as a test bed, and analyze the effect of these parameters on the learned cross-modal representation during training. In addition, we also assess the parameter k that influences the minimax game as described in Algorithm. 1. Specially, we set the range of α, β as $\{0.01, 0.1, 0, 1, 10, 100\}$ and of k as $\{1, 2, 3, 4, 5, 6\}$.



(a) Inter-modal invariance preserved without adversary (b) Inter-modal invariance preserved with adversary (c) Inter-modal invariance and intra-modal discriminativeness preserved with adversary

Figure 4: t-SNE visualization for the test data in the Wikipedia Dataset. The red color represents the visual features and the blue color represents the text features.

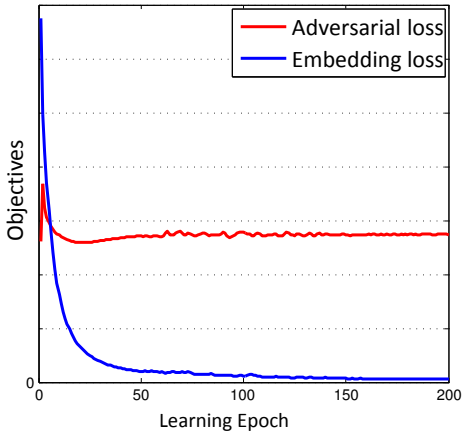


Figure 5: Development of the embedding loss and adversarial loss during the training process, as computed for ACMR on the Wikipedia dataset.

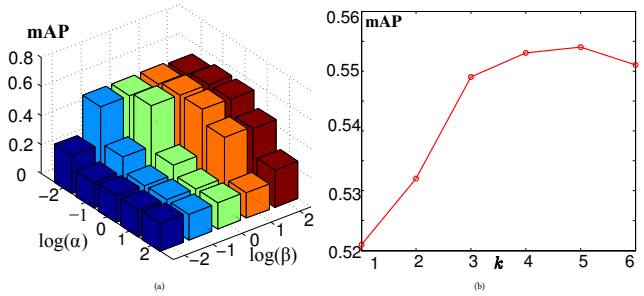


Figure 6: Cross-modal retrieval performance of ACMR with different values of model parameters: (a) α and β; (b) k on Wikipedia dataset.

Note that $\alpha = 0$ and $\beta = 0$ represent the ACMR with $\mathcal{L}_{imi}$ only and ACMR with $\mathcal{L}_{imd}$ only, respectively. The evaluation is conducted by changing one parameter (*e.g.*, α) while fixing the other (*e.g.*, β). Figure 6(a) shows the performance of ACMR for different values of α and β. We can observe that ACMR performs well when α and β are in the range of $[0.01, 0.1]$. Furthermore, the mAP scores obtained by ACMR with $\mathcal{L}_{imi}$ only and ACMR with $\mathcal{L}_{imd}$ only indicate that L_{imd} has more contribution to the overall performance compared to $\mathcal{L}_{imi}$, which is in accordance with the previous observation as shown in Table 5. Figure 6(b) shows the performance of ACMR for different values of k.The figure indicates that, in practice, a dedicated effort to find a proper value of k (*e.g.*, k = 4 or 5) helps the overall optimization process.

5 CONCLUSION

In this paper, we proposed a new approach (ACMR) to learn representations which are both discriminative and modality-invariant for cross-modal retrieval. ACMR is based on an adversarial learning approach that engages two processes in a minimax game: a feature projector that generates modality-invariant and discriminative representations and a modality classifier that tries to detect the modality of an item given an unknown feature representation. We also introduced triplet constraints to ensure that cross-modal semantic data structure is well preserved when projected into common subspace. Comprehensive experimental results on four cross-modal datasets and extensive analysis have demonstrated the effectiveness of our algorithmic and methodological design choices, leading to superior cross-modal retrieval performance compared to state-of-the-art methods. An interesting open issue for future research is to further adjust our proposed ACMR framework to better deal with high cost of training of its DNN architecture.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Project 61572108, Project 61632007, Project 61602089 and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007 and Project ZYGX2015J055.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. 2013. Deep canonical correlation analysis.. In *ICML*. 1247–1255.

[2] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger. 2017. Adversarial deep averaging networks for cross-lingual sentiment classification. (2017). arXiv:1406.2661v1

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. 2009. NUS-WIDE: A real-world web image database from National University of Singapore. In *CIVR*.

[4] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI* 36, 3 (2014), 521–535.

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

[6] Y. Ganin and V. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation.. In *ICML*. 1180–1189.

[7] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV* 106, 2 (2014), 210–233.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[9] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16, 12 (2004), 2639–2664.

[10] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *ICME*.

[11] R. Hong, Y. Yang, M. Wang, and X.-S. Hua. 2015. Learning visual semantic relationships for efficient visual retrieval. *TBD* 1, 4 (2015), 152–161.

[12] A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 3128–3137.

[13] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.

[14] B. Klein, G. Lev, G. Sadeh, and L. Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*. 4437–4446.

[15] D. Li, N. Dimitrova, M. Li, and I.K. Sethi. 2003. Multimedia content processing through cross-modal association. In *ACM MM*. 604–611.

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

[17] Z.-C. Lipton and S. Tripathi. 2017. Precise recovery of latent vectors from generative adversarial networks. In *ICLR Workshop*.

[18] L. Ma, Z. Lu, L. Shang, and H. Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*. 2380–7504.

[19] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.

[21] Y. Peng, X. Huang, and J. Qi. 2016. Cross-media shared representation by hierarchical learning with multiple deep networks. In *IJCAI*. 3846–3853.

[22] Y. Peng, J. Qi, X. Huang, and Yuan Y. 2017. CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. (2017). arXiv:1704.02116

[23] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

[24] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao. 2016. A fast optimization method for general binary code learning. *TIP* 25, 12 (2016), 5610–5621.

[25] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y. Yuan, and H. T. Shen. 2016. Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval. *TCYB* PP (2016), 1–14.

[26] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-Scale image recognition. *CoRR* abs/1409.1556 (2014).

[27] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD*. 785–796.

[28] N. Srivastava and R. Salakhutdinov. 2012. Learning representations for multimodal data with deep belief nets. In *ICML Workshop*.

[29] N. Srivastava and R. Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*. 2222–2230.

[30] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. 2011. Joint feature selection and subspace learning for cross-modal retrieval. *TPAMI* 38, 10 (2011), 2010–2023.

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

[32] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. 2017. A comprehensive survey on cross-modal retrieval. arXiv:1607.06215

[33] L. Wang, Y. Li, and S. Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*. 5005–5013.

[34] X. Xu, F. Shen, Yang Y., H. T. Shen, and X. Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *TIP* 26, 5 (2017), 2494–2507.

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

[36] F. Yan and K. Mikolajczyk. 2015. Deep correlation for matching images and text. In *CVPR*. 3441–3450.

[37] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen. 2016. Zero-shot hashing via transferring supervised knowledge. In *ACM MM*. 1286–1295.

[38] Y. Yang, Z. Zha, Y. Gao, X. Zhu, and T.-S. Chua. 2014. Exploiting web images for semantic video indexing via robust sample-specific loss. *TMM* 16, 6 (2014), 1677–1689.

[39] T. Yao, T. Mei, and C.-W. Ngo. 2015. Learning query and image similarities with ranking canonical correlation analysis. In *ICCV*. 28–36.

[40] X. Zhai, Y. Peng, and J. Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *TCSVT* 24 (2014), 965–978.

[41] Y. Zhang, R. Barzilay, and T. Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. (2017). arXiv:1701.00188

[42] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *ACM MM*. 143–152.

[43] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. 2013. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*. 1070–1076.