


*Article*

# Adversarial Learning Based Semantic Correlation Representation for Cross-Modal Retrieval

Lei Zhu <sup>1,2</sup> , Jiayu Song <sup>1</sup>, Xiangxiang Wei <sup>1,2</sup>, and Ju Long <sup>1,2,\*</sup>

<sup>1</sup> School of Computer Science and Engineering, Central South University, Changsha, 410083, PR China; leizhu@csu.edu.cn (L.Z.); jiayusong@csu.edu.cn (J.S.); xiangxiangwei@csu.edu.cn (X.W.);

<sup>2</sup> Network Resources Management and Trust Evaluation Key Laboratory of Hunan Province, Changsha, 410083, PR China

\* Correspondence: jlong@csu.edu.cn;

**Abstract:** With the rapid development of Internet and the widely usage of smart devices, massive multimedia data are generated, collected, stored and shared on the Internet. This trend makes cross-modal retrieval problem become a hot issue in this years. Many existing works pay attentions on correlation learning to generate a common subspace for cross-modal correlation measurement, and others uses adversarial learning technique to abate the heterogeneity of multi-modal data. However, very few works combine correlation learning and adversarial learning to bridge the inter-modal semantic gap and diminish cross-modal heterogeneity. This paper propose a novel cross-modal retrieval method, named ALSCOR, which is an end-to-end framework to integrate cross-modal representation learning, correlation learning and adversarial. CCA model, accompanied by two representation model, VisNet and TxtNet is proposed to capture non-linear correlation. Beside, intra-modal classifier and modality classifier are used to learn intra-modal discrimination and minimize the inter-modal heterogeneity. Comprehensive experiments are conducted on three benchmark datasets. The results demonstrate that the proposed ALSCOR has better performance than the state-of-the-arts.

**Keywords:** Cross-modal retrieval; Adversarial learning; Semantic correlation; Deep learning;

## 1. Introduction

With the rapid development of Internet and the widely usage of smart devices, huge amounts of multimedia data with various modalities, such as images, texts, videos and audios, etc. are generated, collected, stored and shared on the Internet, as shown in Fig. 1. For example, multimedia sharing services such as Flickr, Pinterest, YouTube shares massive images and videos with textual descriptions. Online encyclopedia such as Wikipedia and Baidu baike stores a tremendous amount of items with texts, images covering knowledges in various fields. Online Social networks, such as Twitter, Facebook and Sina Weibo, provide platforms for users to share their lives by millions of tweets and posts with texts, images or short videos. Other applications on mobile platform, such as Instagram and Douyin, make it possible to share pictures and short videos anytime and anywhere. These multi-modal data are usually used to describe the same events, scenes or objects in our daily life, and users always have the need to search relative multimedia data by the queries of different modalities. This retrieval paradigm is called cross-modal retrieval [1–3], which attracts more and more attentions in the community of multimedia.

In the last decade, lots of approaches have been proposed to address the problem of cross-modal retrieval. The main challenge focused by many researchers is to learn a common subspace in which the representations or embeddings of different modalities can be measured via distance function. Canonical Correlation Analysis (CCA) [4] is a widely used statistical method, which is employed by Rasiwasia et al. [5] to find the correlations between representations of different modalities to learn the common subspace. Following [5], several CCA based researches, such as [6–9] have been presented to



Figure 1. Some examples of multi-modal data from Twitter, YouTube and Flickr.

support cross-modal retrieval. Inspired by deep neural networks (DNN) that play an important role of multimedia analysis and pattern recognition, a number of researchers [10–14] exploit DNN to improve the performance of retrieval by learning the non-linear correlations between modalities.

**Motivation.** The previous cross-modal retrieval approaches aims to learn a common semantic subspace in which the representations of different modalities can be measured easily. However, very few works combine correlation learning and adversarial learning to bridge the inter-modal semantic gap and diminish cross-modal heterogeneity together. To overcome this challenge, for the first time, this paper proposes to combine cross-modal correlation learning and adversarial learning and develop an end-to-end framework to learn bridge the semantic gap and diminish the cross-modal heterogeneity. Different from the existing studies [15–18], we combine deep CCA based cross-modal correlation learning and adversarial learning to not only learn the semantic correlations to bridge the semantic gap between different modalities, but implement a better cross-modal distribution alignment to diminish the cross-modal heterogeneity.

**Our Method.** We propose a novel cross-modal retrieval approach, called Adversarial Learning based Semantic CORrelation Representation (**ALSCOR**). It is a combination of cross-modal correlation learning and adversarial learning. For the cross-modal correlation learning, inspired by deep CCA technique, we design a cross-modal deep representation CCA model which consists of a two-branches network, VisNet and TxtNet. The VisNet is a CNN based model that recieves image samples and maps them into deep representations. The TxtNet is realized by word2vec model, BiLSTM model and a text convolution network, which aims to learn deep representations of texts. The CCA model accompanied by these two model is used to learn the inter-modal correlation. Besides, an intra-modal classifier is used to learn the intra-modal discriminative information. In addition, inspired by generative adversarial network, a modality classifier is utilized to diminish the cross-modal heterogeneity, which is realized by discriminating representations of different modalities.

**Contributions.** The main contributions of this paper can be summarized as follows:

- We formalized the definition of cross-modal retrieval, and propose a novel framework that is a combination of cross-modal correlation learning and adversarial learning. To the best of our

knowledge, this work is the first time to improve the retrieval performance by using deep CCA, BiSLTM, CNN and adversarial learning together.

- To learn the inter-modal non-linear correlation, a two-branches cross-modal correlation model is developed, which is a integration of VisNet, TxtNet and CCA. The VisNet realized by CNN is to generate deep visual representations, and TxtNet is implemented by word2vec, BiLSTM and CNN to learn deep textual representations. An intra-modal classifier is utilized to learn the intra-modal discrimination, and the modality classifier plays the discriminator to diminish the cross-modal heterogeneity in an adversarial manner.
- Comprehensive experiments on three benchmark datasets are conducted. We compare the proposed method with 8 state-of-the-arts. The results demonstrate that our method has great performance for cross-modal retrieval.

**Roadmap.** The remainder of this paper is organized as follows: the related works are reviewed in Section 2. In Section 3, we give the definition of cross-modal retrieval and the related techniques, including cross-modal similarity measurement, deep CCA and GAN. In Section 4, the framework of ALSCOR is introduced in details, including the architecture and the loss. Our experimental results are presented in Section 5, and finally we draw the conclusion in Section 6.

## 2. Related Work

In this section, we review existing studies concerning cross-modal retrieval and deep learning, which are relative to our study. To the best of our knowledge, this work is the first to combine deep CCA method and adversarial learning technique to overcome cross-modal retrieval task.

### 2.1. Cross-Modal Retrieval

Cross-modal retrieval is a significant problem in the area of multimedia computing [19–26], which aims to find out the similar enough objects of one modality in the multimedia database by a query of different modality. Due to the exponential growth of amount of multimedia data, this task attracts a large number of attentions in recent years. CCA [27] is an important statistic method to seek the linear correlation between two sets of variates, which is utilized by many studies for cross-modal retrieval. For example, [5] is the first work using CCA to address cross-modal retrieval. In this work, Rasiwasia et al. modeled images and texts by SIFT features and hidden topic model respectively, and then maps the cross-modal representations into a common subspace by CCA. Wang et al. [6] proposed a method called Unsupervised Discriminant Canonical Correlation Analysis (UDCCA), which utilizes normalized spectral clustering to compute class membership. Zu et al. [7] proposed a novel approach named Canonical Sparse Cross-view Correlation Analysis (CSCCA) to consider structure and cross view information. Gong et al. [8] presented a three-view CCA approach that incorporates a third view to model semantic information to improve the retrieval performance. Zhang et al. [9] proposed a method named mixture of probabilistic CCA (MixPCCA) to model the nonlinear correlations between different modalities. Shao et al. [28] presented hypergraph semantic embedding (HSE) approach to model latent semantics from text to regularize the deep CCA subspace. Wang et al. [29] developed a novel correlation subspace learning method by integrating structured sparsity regularization and intra-modal information to achieve better performance. For cross-modal image clustering problem, Jin et al. [30] proposed a CCA based multimodal feature fusion method to characterize the multimodal correlations between the visual features in images and semantic features in captions. Different from the existing researches, we combine deep CCA and adversarial learning method to learn better representations with modality invariance, which can implement feature distribution alignment between different modalities.

Latent Dirichlet Allocation [31] (LDA) is another classical method for text feature representations. Lots of researches use it to extract semantic information to support multi-modal/cross-modal retrieval. Yakhnenko [32] introduced a LDA based method called multi-modal hierarchical Dirichlet Process (MoM-HDP) to model multi-modal data. Putthividhya et al. [33] proposed topic-regression

multi-modal Latent Dirichlet Allocation (tr-mmLDA) method to capture correlations between cross-modal data. Blei et al. [34] presented correspondence latent Dirichlet allocation method to learn the joint distribution of multi-modal data. Jia et al. [35] proposed a method named Multi-modal Document Random Field to model the relations between different modalities. Lu et al. [36] model the cross-modal retrieval problem as a listwise ranking problem, and proposed a method named Latent Semantic Cross-Modal Ranking (LSCMR) to learn a latent space. Yu et al. [37] utilized LDA to model texts, which is to learn latent semantic relations between texts and images. Zoghbi et al. [38] used bilingual LDA and CCA to model textual and visual data in the task of cross-modal attribute recognition. Instead of LDA, in this work we utilize biLSTM and CNN model to generate deep representations of texts, which is to capture more high-level semantic features.

## 2.2. Deep Learning

As a very powerful technique, deep learning [39–41] is used to overcome several challenges of multimedia analysis and retrieval, such as image classification [42], object recognition [43], video retrieval [44], multi-modal/cross-modal retrieval [45], etc. Wei et al. [46] proposed to use deep CNN to learn visual representations for cross-modal retrieval, which performs much better than traditional hand-crafted features. He et al. [47] employed convolution-based networks to generate both visual and textual representations. Shen et al. [48] introduced a method, Textual-Visual Deep Binaries (TVDB), to encode semantics of informative images and long textual descriptions. Yang et al. [49] presented an end-to-end deep learning architecture to generate compact cross-modal hash codes from intra-modal and an inter-modal view. Cao et al. [50] developed a collective deep quantization (CDQ) method, which is an end-to-end deep architecture to jointly learn deep cross-modal representation and quantizers. Hu et al. [51] presented a method named Dense Multimodal Fusion (DMF) to generate joint representations hierarchically, which can learn the correlations in different levels. Gu et al. [52] proposed to use generative model to learn not only the global abstract features but also the local grounded features.

Inspired by [53], several studies utilized adversarial learning to improve the cross-modal representation learning. Wang et al. [15] is the first to use adversarial learning method to combat cross-modal retrieval problem. Zhang et al. [16] proposed to adversarial learning based method to learn attention mask for cross-modal feature generation. Wen et al. [17] introduced a new cross-modal similarity transferring (CMST) method by adversarial learning. This method is to learn common representation subspace via quantitative similarities in single-modal representation subspace. Shang et al. [18] developed a dictionary learning based adversarial cross-modal retrieval technique, which makes the transformed features maintain the inherent statistical characteristics of original features. Unlike the existing studies, we combine deep CCA method and adversarial learning to not only learn the semantic correlations between different modalities, but implement a better cross-modal distribution alignment.

## 3. Preliminary

In this section, we firstly formalize the definition of cross-modal retrieval and the related notions. Besides, the cross-modal correlation measurement is proposed. In addition, two important techniques, i.e., deep CCA and LSTM are introduced, which are related to our method. Table 1 summarizes the notations frequently used in this paper to facilitate the discussion.

### 3.1. Problem Definition

**Definition 1 (Cross-Modal Retrieval).** Without losing generality, consider a multimedia dataset that contains multi-modal data, is denoted as  $\mathcal{D} = \{X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_n\}$ , where  $X$  and  $Y$  denote two



Notation	Definition
$\mathcal{D}$	A multimedia dataset
$I$	An image
$T$	A text
$\mathcal{R}$	A result set
$\mathbf{L}_i$	A classification label vector of $i$ -th image-text pair.
$L_i^{(c)}$	The $c$ -th classification label of $i$ -th image-text pair.
$\text{Corr}(\cdot, \cdot)$	The correlation measurement function
$\Xi$	The representation of image
$\zeta$	The representation of text
$\mathbf{M}_I$	The image representation matrix
$\mathbf{M}_T$	The text representation matrix
$\mathbf{M}_L$	The classification label representation matrix
$\Omega_I$	The image mapping
$\Omega_T$	The text mapping
$\theta_I$	The model parameter vector of VisNet
$\theta_T$	The model parameter vector of TxtNet
$\theta_D$	The model parameter vector of intra-modal classifier
$\theta_A$	The model parameter vector of modality discriminator
$\bowtie$	The vector concatenation operator
$\mathbf{K}^{(\kappa)}$	The $\kappa$ -th convolutional kernel
$f_i^{(j)}(t)$	The $j$ -th location of the input map of $i$ -th samples at time $t$

**Table 1.** The summary of notations that are frequently used in this paper

different modalities. Cross-modal retrieval aims to return a set of data  $\mathcal{R}$  of one modality, which are correlative enough to the query of another modality, namely,

$$\mathcal{R} = \{Y | Y \in \mathcal{D}, Y' \in \mathcal{D} \setminus \mathcal{R}, \text{Corr}(X_Q, Y) \geq \text{Corr}(X_Q, Y')\} \quad (1)$$

where  $\text{Corr}(\cdot, \cdot)$  is a cross-modal correlation measurement that is to measure the correlations between two objects of different modalities.

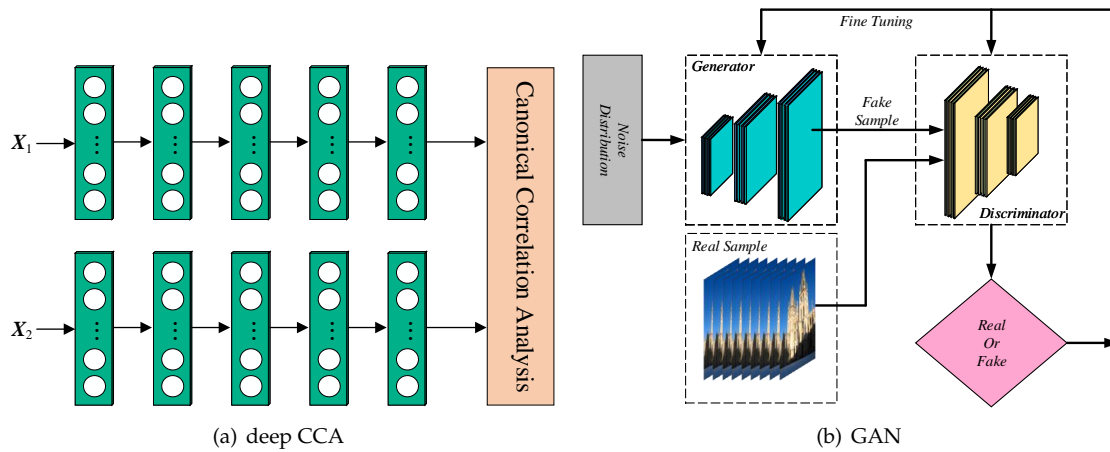
This work focuses on two most common modalities on the Internet, i.e., image  $I$  and text  $T$ . Two corresponding retrieval tasks are studied: (1) image-to-text retrieval that is to find correlative texts by a image query and (2) text-to-image retrieval that is to find correlative images by a text query. According to Definition 1, these two tasks can be formalized as

$$\mathcal{R}_{I2T} = \{T | T \in \mathcal{D}, T' \in \mathcal{D} \setminus \mathcal{R}_{I2T}, \text{Corr}(I_Q, T) \geq \text{Corr}(I_Q, T')\}, \quad (2)$$

$$\mathcal{R}_{T2I} = \{I | I \in \mathcal{D}, I' \in \mathcal{D} \setminus \mathcal{R}_{T2I}, \text{Corr}(T_Q, I) \geq \text{Corr}(T_Q, I')\}. \quad (3)$$

Suppose that the multimedia dataset  $\mathcal{D} = \{\langle I_1, T_1 \rangle, \langle I_2, T_2 \rangle, \dots, \langle I_n, T_n \rangle\}$  contains  $n$  pairs of image and text. Each pairs has a classification label vector denoted as  $\mathbf{L}_i = \{L_i^{(1)}, L_i^{(2)}, \dots, L_i^{(c)}\} \in \mathbb{R}^c$ , where  $c$  is the number of the classifications. If the  $i$ -th object belongs to the  $j$ -th classification, the  $L_i^{(j)} = 1$ , otherwise  $L_i^{(j)} = 0$ . The representations of image  $I_i$  and text  $T_i$  are denoted as  $\xi_i = (\xi_i^{(1)}, \xi_i^{(2)}, \dots, \xi_i^{(\gamma_I)}) \in \mathbb{R}^{\gamma_I}$  and  $\zeta_i = (\zeta_i^{(1)}, \zeta_i^{(2)}, \dots, \zeta_i^{(\gamma_T)}) \in \mathbb{R}^{\gamma_T}$ , where  $\gamma_I$  and  $\gamma_T$  are the number of demensions, and generally  $\gamma_I \neq \gamma_T$ . Therefore, for the multimedia dataset  $\mathcal{D}$ , the image representation matrix, the text representation matrix and the classification label matrix are denoted as  $\mathbf{M}_I = (\xi_1, \xi_2, \dots, \xi_n) \in \mathbb{R}^{\gamma_I \times n}$ ,  $\mathbf{M}_T = (\zeta_1, \zeta_2, \dots, \zeta_n) \in \mathbb{R}^{\gamma_T \times n}$ , and  $\mathbf{M}_L = (\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_n) \in \mathbb{R}^{c \times n}$ , respectively.

The main challenge of implementing cross-modal retrieval is the heterogeneity between different modalities, which is manifested in two aspects: (1) the difference of feature distributions and (2) the semantic gap between different modalities. That means the demensions of feature representations are different, and they are hard to be represented in the same distribution. Besides, the semantic concepts



**Figure 2.** The structure of deep CCA and GAN.

of representations are hard to be aligned. These two limitations hinder the correlation measurement of cross-modal data. Thus, two cross-modal mappings,  $\Omega_I((I_1, I_2, \dots, I_n))$  and  $\Omega_T((T_1, T_2, \dots, T_n))$  need to be learnt to project images  $(I_1, I_2, \dots, I_n)$  and texts  $(T_1, T_2, \dots, T_n)$  into a common semantic subspace, in which the representations of images and texts have the similar distributions and the semantic concepts can be aligned. Formally,

$$\Omega_I((I_1, I_2, \dots, I_n)) : (I_1, I_2, \dots, I_n) \rightarrow (\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_n) \in \mathbb{R}^{\gamma \times n}, \quad (4)$$

$$\Omega_T((T_1, T_2, \dots, T_n)) : (T_1, T_2, \dots, T_n) \rightarrow (\hat{\zeta}_1, \hat{\zeta}_2, \dots, \hat{\zeta}_n) \in \mathbb{R}^{\gamma \times n}, \quad (5)$$

where  $(\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_n)$  and  $(\hat{\zeta}_1, \hat{\zeta}_2, \dots, \hat{\zeta}_n)$  are the representation matrix of images and texts in the common semantic subspace. Therefore, the correlations between multi-modal data can be measured via distance function in this space. Inspired by Pearson correlation, we propose the cross-modal correlation measurement as follows:

**Definition 2 (Cross-Modal Correlation Measurement).** Given an image  $I_i$  and a text  $T_j$ , the representations of  $I_i$  and  $T_j$  in the common semantic subspace are  $\hat{\xi}_i$  and  $\hat{\zeta}_j$ , the cross-modal correlation between  $I_i$  and  $T_j$  is measured by the following equation:

$$\text{Corr}(I_i, T_j) = \frac{\sum_{k=1}^{\gamma} (\hat{\xi}_i^{(k)} - \mu_{\hat{\xi}_i}) \times (\hat{\zeta}_j^{(k)} - \mu_{\hat{\zeta}_j})}{\sqrt{\sum_{k=1}^{\gamma} (\hat{\xi}_i^{(k)} - \mu_{\hat{\xi}_i})^2} \times \sqrt{\sum_{k=1}^{\gamma} (\hat{\zeta}_j^{(k)} - \mu_{\hat{\zeta}_j})^2}} \quad (6)$$

where  $\mu_{\hat{\xi}_i}$  and  $\mu_{\hat{\zeta}_j}$  are the averages of  $\hat{\xi}_i$  and  $\hat{\zeta}_j$ , respectively.

### 3.2. Deep Canonical Correlation Analysis

Deep CCA is an extension proposed by Andrew et al. [54], which is a combination of linear CCA and deep neural network to learn the non-linear correlation between two views. As shown in Fig. 2 (a), the deep CCA consists of a coupled  $d$ -layer fully-connected neural networks. The input instances of view 1 and view 2 are denoted as  $X_1 \in \mathbb{R}^{m_1}$  and  $X_2 \in \mathbb{R}^{m_2}$ , respectively. The output of the first neural network layer are  $h_1^1 = s(W_1^1 X_1 + b_1^1) \in \mathbb{R}^{c_1}$  and  $h_2^1 = \Sigma(W_2^1 X_2 + b_2^1) \in \mathbb{R}^{c_2}$ , where  $W_1^1 \in \mathbb{R}^{c_1 \times m_1}$  and  $W_2^1 \in \mathbb{R}^{c_2 \times m_2}$  are the matrices of weight,  $b_1^1$  and  $b_2^1$  are the vectors of bias,  $\Sigma(\cdot)$  is a non-linear activation function. The second layer receives the outputs of the first layer and generates its outputs  $h_1^2 = \Sigma(W_1^2 h_1^1 + b_1^2) \in \mathbb{R}^{c_1}$  and  $h_2^2 = \Sigma(W_2^2 X_2 + b_2^2) \in \mathbb{R}^{c_2}$ . For the last layer, the outputs are  $f_1(X_1) = \Sigma(W_1^d h_1^{d-1} + b_1^{d-1}) \in \mathbb{R}^o$  and  $f_2(X_2) = \Sigma(W_2^d h_2^{d-1} + b_2^{d-1}) \in \mathbb{R}^o$ . The objective of the deep CCA is to learn the parameters of the two-way networks to maximize the correlation between  $X_1$  and

$X_2$ , namely  $\text{Corr}(f_1(X_1), X_2)$ . Let  $\theta_1 = (W_1^1, \dots, W_1^d, b_1^1, \dots, b_1^d)$  and  $\theta_2 = (W_2^1, \dots, W_2^d, b_2^1, \dots, b_2^d)$  be the network parameter vectors of view 1 and view 2, the objective function is:

$$(\theta_1^*, \theta_2^*) = \arg \max_{(\theta_1, \theta_2)} \text{Corr}(f_1(X_1; \theta_1), f_2(X_2; \theta_2)). \quad (7)$$

This objective function can be optimized by gradient descent method on the training set. Let  $H_1 \in \mathbb{R}^{o \times n}$  and  $H_2 \in \mathbb{R}^{o \times n}$  be the representations matrices generated by top-layer networks, where  $n$  is the number of training samples. Let  $\bar{H}_1 = H_1 - \frac{1}{n}H_1\mathbf{1}$  and  $\bar{H}_2 = H_2 - \frac{1}{n}H_2\mathbf{1}$  be the centered matrices,  $\hat{\Sigma}_{11} = \frac{1}{n-1}\bar{H}_1\bar{H}_1' + r_1I$ ,  $\hat{\Sigma}_{22} = \frac{1}{n-1}\bar{H}_2\bar{H}_2' + r_2I$ ,  $\hat{\Sigma}_{12} = \frac{1}{n-1}\bar{H}_1\bar{H}_2'$ , where  $\mathbf{1} \in \mathbb{R}^{o \times o}$  is an all-1 matrix,  $I$  is an identity matrix,  $r_1$  and  $r_2$  are the regularization terms. Here suppose that  $r_1 > 0$  and  $r_2 > 0$ , thus both of  $\hat{\Sigma}_{11}$  and  $\hat{\Sigma}_{22}$  are positive definite. As the total correlation of the top- $k$  components of matrices  $H_1$  and  $H_2$  is the sum of the top- $k$  singular values of  $T = \hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}$ , let  $k = o$ , the correlation between  $H_1$  and  $H_2$  can be computed by the following equation:

$$\text{Corr}(H_1, H_2) = \text{tr}(T'T)^{-1/2}. \quad (8)$$

Let the singular value decomposition of matrix  $T$  be  $T = UDV'$ , then

$$\frac{\partial \text{Corr}(H_1, H_2)}{\partial H_1} = \frac{1}{n-1}(2\nabla_{11}\bar{H}_1 + \nabla_{12}\bar{H}_2), \quad (9)$$

$$\frac{\partial \text{Corr}(H_1, H_2)}{\partial H_2} = \frac{1}{n-1}(2\nabla_{22}\bar{H}_2 + \nabla_{12}\bar{H}_1), \quad (10)$$

where

$$\nabla_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}U D U' \hat{\Sigma}_{11}^{-1/2}, \quad (11)$$

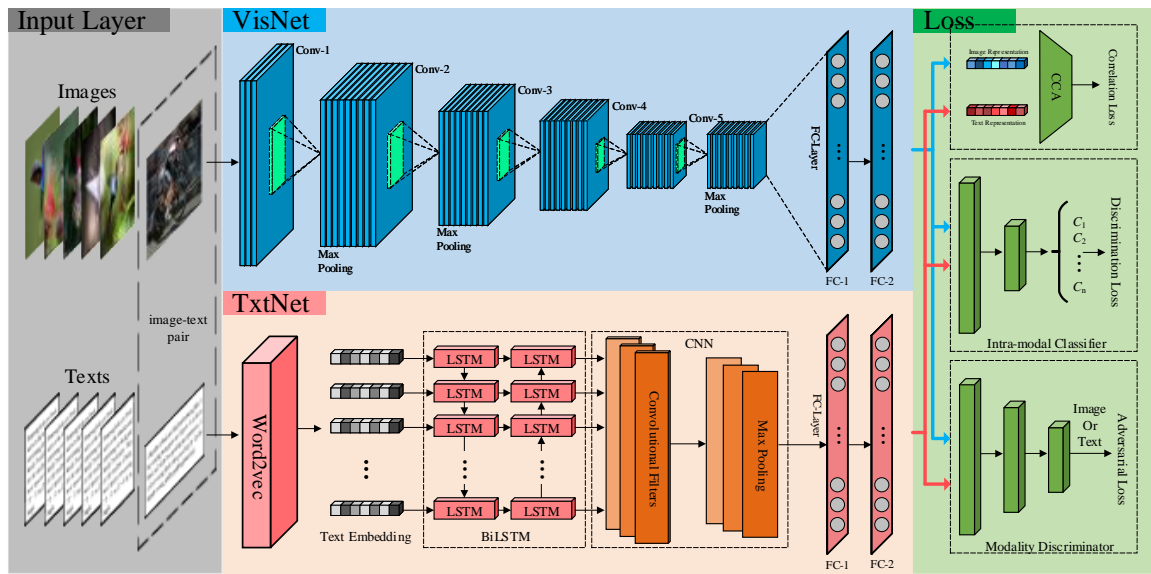
$$\nabla_{22} = -\frac{1}{2}\hat{\Sigma}_{22}^{-1/2}U D U' \hat{\Sigma}_{22}^{-1/2}, \quad (12)$$

$$\nabla_{12} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2}U V' \hat{\Sigma}_{22}^{-1/2}. \quad (13)$$

### 3.3. Generative Adversarial Network

Generative Adversarial Network (GAN) [53] is an unsupervised learning approach, which is used for several multimedia tasks, such as image generation, 3D reconstruction, super-resolution, etc. As shown in Fig. 2 (b), GAN consists of two component: a generator  $G$  and a discriminator  $D$ . The generator aims to synthesize forged image according to real samples, and the discriminator is to recognize whether the inputs are the productions of generator or the natural images. The training process is equivalent to a two-player zero-sum game: the generator  $G$  endeavours to produce synthetic samples more similar to the real samples, while the discriminator  $D$  vigorously to identify whether the inputs is from the distribution of  $G$  or the natural distribution. At last, these two models achieve a dynamic equilibrium: the generated samples are similar enough to the natural distribution, while the discriminator  $D$  cannot discriminate the real and synthetic samples. Let a natural image sample be  $I$  that obey natural random distribution  $P(I)$ ,  $z \in \mathbb{R}^\gamma$  be a random vector from distribution  $P_z(z)$ . The generator  $G(I; \theta_G)$  aims to map  $z$  to a synthetic sample  $G(z; \theta_G)$ . The discriminator  $D(I; \theta_D)$  receives  $G(z; \theta_G)$  as input and outputs the discriminant result  $D(G(z; \theta_G); \theta_D)$  which is the probability that  $G(z; \theta_G)$  is generated from  $G$ . Thus, this game process can be formulated as an minimax optimization, formalized by the following objective function:

$$\arg \min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}_{I \sim P_n(I)} [\log D(I; \theta_D)] + \mathbb{E}_{z \sim P_z(z)} [\log(1 - D(G(z; \theta_G); \theta_D))] \quad (14)$$



**Figure 3.** The architecture of the proposed ALSCOR. ALSCOR is an end-to-end framework of cross-modal retrieval. It has three layers: (1) Input layer that feeds image-text pairs into the cross-modal representation layer; (2) cross-modal Representation layer includes two models: VisNet and TtxtNet; (3) Loss layer contains three models. CCA model is used to learn the correlations between representations of image and text via correlation loss. Intra-modal Classifier model is to learn the intra-modal discriminative representations via discrimination loss. The modality discriminator is used to learn modality-invariant representations via adversarial loss.

where  $\theta_G$  and  $\theta_D$  are the model parameters of  $G$  and  $D$ , respectively.  $\mathbb{E}_{I \sim P_n(I)}[\cdot]$  and  $\mathbb{E}_{z \sim P_z(z)}[\cdot]$  are the expectations, and

$$\mathbb{E}_{I \sim P_n(I)}[\log D(I; \theta_D)] = \int_I P_n(I) \log(D(I; \theta_D)) dI, \quad (15)$$

$$\mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z; \theta_G); \theta_D))] = \int_z P_z(z) \log(1 - D(G(z; \theta_G); \theta_D)) dz. \quad (16)$$

The generator  $G$  and discriminator  $D$  are trained in an alternate and iterative manner. For the generator  $G$ , the objective is to minimize the loss function to produce more authentic samples to fool the discriminator  $D$ . By contrast, for the discriminator  $D$ , the objective is to maximize the loss. Formally, the objective functions are:

$$\arg \min_G \mathcal{L}_{GAN}(G, D) = \int_I P_n(I) \log(D(I; \theta_D)) dI \quad (17)$$

$$\arg \max_D \mathcal{L}_{GAN}(G, D) = \int_z P_z(z) \log(1 - D(G(z; \theta_G); \theta_D)) dz \quad (18)$$

#### 4. The Method

To learn a common subspace and bridge the semantic gap between different modalities, we propose an effective end-to-end framework, named **Adversarial Learning based Semantic CORrelation Representation (ALSCOR)**, which is a combination of cross-modal correlation learning and adversarial learning. In this section, we introduce this method in details.

##### 4.1. The Architecture of ALSCOR

**Overview.** The general architecture of the proposed ALSCOR is illustrated in Fig. 3, which is a combination of deep CCA and adversarial learning technique. However, different from the



traditional deep CCA that uses two deep fully connected networks, ALSCOR utilizes a deep CNN for image representation and a integration of BiLSTM and convolutional network for text representations. Specifically, this is an end-to-end framework of cross-modal retrieval. It has three layers: (1) **Input layer**, (2) **cross-modal Representation layer** and (3) **Loss layer**. The input layer contains a multimedia collection, and it feeds image-text paris into the next layer, namely cross-modal representation layer. Cross-modal Representation layer is a two-way deep neural network structure, which includes two models to learn deep representations of image and text, respectively. The one is VisNet, which is implemented by deep convolutional network to learn the deep representations of images. The other is TxtNet, which is generate text representations via a comnination of word2vec model, BiLSTM network and text convolutional network. The Loss layer contains three models. CCA model is used to learn the correlations between representations of image and text via correlation loss. Intra-modal Classifier model is to learn the intra-modal discriminative representations via discrimination loss. The modality discriminator is used to learn modality-invariant representations via adversarial loss.

**VisNet.** VisNet is a CNN based model to generate deep representations of images, formally,  $(\zeta_i^{(1)}, \zeta_i^{(2)}, \dots, \zeta_i^{(\gamma_I)}) = \text{VisNet}(I_i; \theta_I)$ , where  $\theta_I$  is the model parameter vector. Compared with deep fully-connected neural networks in deep CCA method, CNN is more powerful to capture high-level visual semantic information from images. In our approach, we utilize AlexNet [42], a well-known CNN model to implement VisNet. Specifically, it has five convolutional layers and two fully-connected layers. The input images are resized to  $224 \times 224 \times 3$ , which are fed into the first convolutional layer. The first convolutional layer has 96 kernels of size  $11 \times 11 \times 3$ . The second convolutional layer has 256 kernels of size  $5 \times 5 \times 96$ . The third convolutional layers has 384 kernels of size  $3 \times 3 \times 256$ . The fourth convolutional layers has 384 kernels of size  $3 \times 3 \times 192$ . The fifth convolutional layers has 256 kernels of size  $3 \times 3 \times 192$ . Following the last convolutional layer, there are two fully-connected layers that have 4096 neurons each. The second full-connected layer output 4096-dimensional feature representations, namely  $\gamma_I = 4096$ .

**TxtNet.** TxtNet is a combination of word2vec, BiLSTM network and CNN model, which is to generate  $\lambda$ -dimensional text representations, namely  $(\zeta_i^{(1)}, \zeta_i^{(2)}, \dots, \zeta_i^{(\gamma_T)}) = \text{TxtNet}(T_i; \theta_T)$ , where  $\theta_T$  is the model parameter vector. The word2vec model receives the inputs and generates a set of word embeddings  $(v_1, v_2, \dots, v_n)$ , where  $n$  is the number of the words. Compared to the traditional representation method (such as BoW), word2vec can capture both semantic and synthactic information of text. Following word2vec model, a BiLSTM model is used to encode the contextual information from both the previous and future context. For each of the directions, the LSTM has three gates: the input gate  $i$ , forget gate  $f$  and output gate  $o$ . At the time  $t$ , the each state in the LSTM is:

$$i(t) = \sigma(W_i[h(t-1), v(t)] + b_i) \quad (19)$$

$$f(t) = \sigma(W_f[h(t-1), v(t)] + b_f) \quad (20)$$

$$o(t) = \sigma(W_o[h(t-1), v(t)] + b_o) \quad (21)$$

$$\tilde{C}(t) = \tanh(W_c[h(t-1), v(t)] + b_c) \quad (22)$$

$$C(t) = i(t) \times \tilde{C}(t) + f(t) \times C(t-1) \quad (23)$$

$$h(t) = o(t) \times \tanh(C(t)) \quad (24)$$

where  $h(t)$  is the hidden vector,  $\sigma$  is the sigmoid function,  $W_i, W_f, W_o, b_i, b_f, b_o$  are the model parameter vectors. The output of the BiLSTM is the concatenation of the outputs of two LSTMs, namely  $h(t) = h_1(t) \bowtie h_2(t)$ , where  $\bowtie$  is the vector concatenation operator.

After the BiLSTM model, a convolutional network is used to capture the local semantic information of the output of BiLSTM. This network has one convolutional layer with  $\kappa$  convolutional kernels of

size  $m \times m$ , i.e.,  $\mathcal{K} = (\mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \dots, \mathbf{K}^{(\kappa)})$ . For the  $j$ -th location of the input map, the calculation can be formalized as follows:

$$f_i^{(j)}(t) = g\left(\sum_{i=1}^m \mathbf{h}(t+i-1)^T * \mathbf{K}^{(i)} + b\right) \quad (25)$$

where  $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$  is a non-linear activation function,  $b$  is a bias,  $*$  is the convolutional operator. For the  $i$ -th kernel  $\mathbf{K}^{(i)}$ , it slides across the input feature map step-by-step and generate the feature map as follows:

$$f_i(t) = (f_i^{(1)}(t), f_i^{(2)}(t), \dots, f_i^{(H-m+1)}(t)) \quad (26)$$

where  $H$  is the size of the input map. All the feature maps are denoted as  $(f_1(t), f_2(t), \dots, f_m(t))$ , which are fed into the max-pooling layer, namely,

$$(\hat{f}_1(t), \hat{f}_2(t), \dots, \hat{f}_m(t)) = \max(f_1(t), f_2(t), \dots, f_m(t)) \quad (27)$$

where  $\max(\cdot)$  is the function to select the maximal element of each feature vector.

To restraint overfitting, before the fully-connected layers, a drop-out operation is used to randomly discards a part of outputs of max-pooling, shown as follows:

$$(\zeta_i^{(1)}, \zeta_i^{(2)}, \dots, \zeta_i^{(\gamma_T)}) = \mathbf{W}_{fc} \times (\hat{f}_1(t), \hat{f}_2(t), \dots, \hat{f}_m(t))^T \odot \mathcal{M} + \beta \quad (28)$$

where  $\mathbf{W}_{fc}$  and  $\beta$  are the parameters of the fully-connected layers,  $\odot$  is the elementwise multiplication operator,  $\mathcal{M}$  is a masking vector of Bernoulli random variables.

#### 4.2. The Loss

In the loss layer, three modules are used to learn the common semantic subspace. The CCA module aims to learn the correlation between images and texts by correlation loss, which receives the deep representations from VisNet and TxtNet. The intra-modal classifier is to learn the intra-modal discriminations by using the classification labels of image-text pairs via discrimination loss. The modality classifier plays an role of discriminator in GAN, which is to diminish the heterogeneity between representations of different modalities via adversarial loss.

**Correlation Loss.** The CCA module integrated with VisNet and TxtNet forms a end-to-end non-linear correlation learning model to maximize the cross-modal correlation. According to deep CCA, the correlation loss is formalized as follows:

$$\mathcal{L}_{corr}(I_i, T_i; \theta_I, \theta_T) = \text{Corr}(\text{VisNet}(I_i; \theta_I), \text{TxtNet}(T_i; \theta_T)). \quad (29)$$

Therefore, the correlation learning is to optimize the following objective function:

$$(\theta_I^*, \theta_T^*) = \arg \max_{(\theta_I, \theta_T)} \mathcal{L}_{corr}(I_i, T_i; \theta_I, \theta_T), \forall \langle I_i, T_i \rangle \in \mathcal{D}. \quad (30)$$

The optimal parameters  $(\theta_I^*, \theta_T^*)$  are calculated by using gradient of the correlation objective on the training set  $\mathcal{D}$ . The optimization can be followed the equation 9 to 13.

**Discrimination Loss.** The intra-modal classifier is to maintain the discrimination of multi-modal data after the cross-modal non-linear mapping. It is realized by predicting the categories label of the cross-modal data in the common semantic subspace. Specifically, this model is a feed-forward neural network followed by a softmax layer, which is to receives the representations of different modalites in the common subspace and output a probability distribution of categories. In our scheme, the cross-entropy loss is used to implement the discrimination loss, shown as follows:

$$\mathcal{L}_{disc}(I_i, T_i; \theta_I, \theta_T, \theta_D) = -\frac{1}{m} \sum_{i=1}^m L_i(\log P(\text{VisNet}(I_i; \theta_I)) + \log P(\text{TxtNet}(T_i; \theta_T))) \quad (31)$$

where  $P(\cdot)$  is the probability distribution of categories,  $\theta_D$  is the parameter vector of the classifier,  $m$  is the number of samples in each mini-batch during the training.

**Adversarial Loss.** Inspired by GAN, the adversarial learning in our method is realized by a modality classifier  $D$ , which works as the discriminator to identify the representation is generated from an image and a text. According to [15], this model is implemented by a three-layer feed-forward neural network with parameter  $\theta_A$ . The representations generated from image modality are assigned with label 01, and the representations generated from text modality are assigned with label 10. The loss function can be formalized as follows:

$$\begin{aligned} \mathcal{L}_{adv}(I_i, T_i; \theta_I, \theta_T, \theta_A) \\ = -\frac{1}{n} \sum_{i=1}^n m_i (\log D(\text{VisNet}(I_i; \theta_I); \theta_A) + \log(1 - D(\text{TxtNet}(T_i; \theta_T); \theta_A))) \end{aligned} \quad (32)$$

where  $m_i$  is the ground-truth modality label of each representation.

**Adversarial Learning.** For the adversarial learning process, we incorporate the correlation loss (equation 29), discrimination loss (equation 31) and adversarial loss (equation 32), and optimize them as a min-max game, shown as follows:

$$(\theta_I^*, \theta_T^*, \theta_D^*) = \arg \min_{(\theta_I, \theta_T, \theta_D)} (\alpha \mathcal{L}_{corr}(I_i, T_i; \theta_I, \theta_T) + \delta \mathcal{L}_{disc}(I_i, T_i; \theta_I, \theta_T, \theta_D) - \varepsilon \mathcal{L}_{adv}(I_i, T_i; \theta_I, \theta_T, \theta_A^*)), \quad (33)$$

$$\theta_A^* = \arg \max_{\theta_A} (\alpha \mathcal{L}_{corr}(I_i, T_i; \theta_I^*, \theta_T^*) + \delta \mathcal{L}_{disc}(I_i, T_i; \theta_I^*, \theta_T^*, \theta_D^*) - \varepsilon \mathcal{L}_{adv}(I_i, T_i; \theta_I, \theta_T, \theta_A)), \quad (34)$$

$$\forall \langle I_i, T_i \rangle \in \mathcal{D}.$$

where  $\alpha$ ,  $\delta$  and  $\varepsilon$  are the weight parameters for these three loss terms. The training can be realized by using a stochastic gradient descent algorithm.

## 5. Experiments

In this section, we present the performance evaluation of the proposed method and the comparison with several state-of-the-arts on four multimedia datasets. The experimental setup is introduced in section 5.1, the implementation details of the proposed method is described in section 5.2, and the performance evaluation and comparison are shown in section 5.3.

### 5.1. Experimental Setup

**Datasets.** Our experiments are conducted on three benchmark multimedia datasets: Wikipedia [5], NUS-WIDE [55], and Pascal Sentence [56], which are widely used in multi-modal/cross-modal retrieval tasks. The detailed descriptions are presented as follows. Some samples of these datasets are shown in Fig. 4.

- **Wikipedia.** Wikipedia (<https://en.wikipedia.org/wiki/>) dataset is generated from the "Wikipedia feature articles". It contains 2866 image-text documents that are divided into ten different semantic categories, including art & architecture, biology, geography & places, history, literature & theater, media, music, royalty & nobility, sport & recreation and warfare. For each document, the text part is the description of the corresponding image content. Following the experimental settings in [5], Wikipedia dataset in our experiments is split randomly into a training set containing size 2173 image-text documents, and a testing set containing 693 documents. To compared with the proposed multi-modal convolutional representations, for each image, the SIFT [57] descriptor is used to produce hand-crafted visual features and then the vocabulary is built by  $k$ -means clustering to produce 1000-dimensional Bag-of-Visual-Words (BoVW) [58] vector. For each text, the 3000-dimensional Bag-of-Words (BoW) vector is generated.

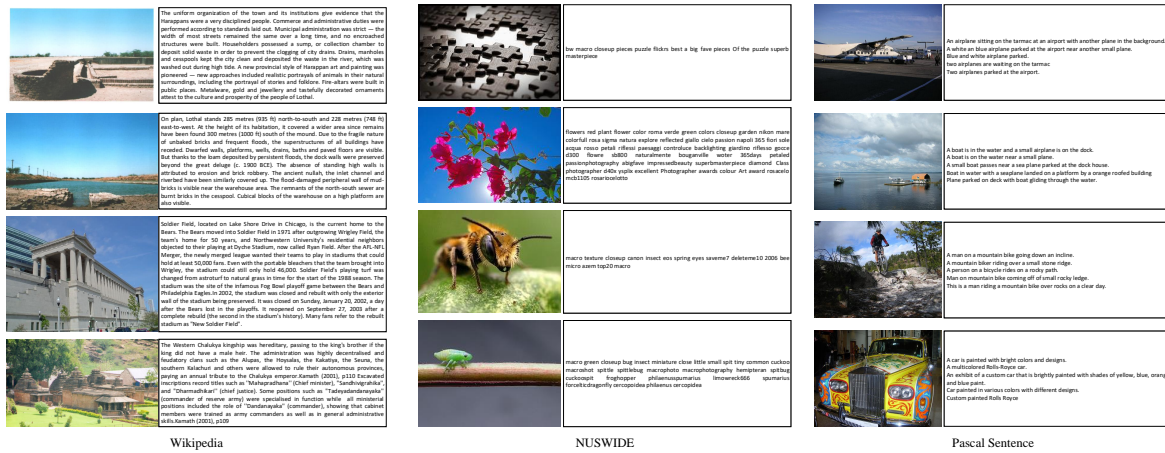


Figure 4. Some image-text pairs in Wikipedia, NUS-WIDE and Pascal Sentence dataset.

- NUS-WIDE.** NUS-WIDE (<https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/NUS-WIDE/NUS-WIDE.html>) dataset is a real-world web image dataset which contains 269648 images and the associated tags from Flickr, with a total number of 5018 unique tags. Each image in NUS-WIDE corresponds to 81 ground truth labels and 1000 text tags. We select the labeled images which belong to the 21 largest categories for experiments. The training set has 114117 samples and the testing set contains 76303 samples. Following the experimental settings in [55], the hand-crafted features of iamges are 500-dimensional SIFT descriptions based BoVW vectors and the textual features are 1000-dimensional BoW vector.
- Pascal Sentence.** Pascal Sentence (<http://vision.cs.uiuc.edu/pascal-sentences/>) dataset aims to support for pattern analysis, statistical modeling and computational learning, which is a subset of Pascal VOC. It contains twenty categories, including aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv. The training set consists of 10103 images with 23374 objects such that there are approximately 500 training objects per category. To compare with CNN representations, we generate 1024-dimensional SIFT based BoVW feature vectors for images, and 100-dimensional topic probability features by LDA and BoW for texts.

**Compared Approaches.** To verify the effectiveness of the proposed method, we compare the proposed ALSCOR with 8 approaches on these three datasets. The brief introduction of these approaches are shown as follows:

- CCA.** CCA [27] is a multivariate statistical analysis method to find the linear correlation between two multivariate random variables. It is widely used in many multi-modal/cross-modal retrieval tasks.
- DCCA.** Deep Canonical Correlation Analysis [54] (DCCA for short) is an extension of CCA to learn complex nonlinear transformations of two representation vectos via two deep neural networks.
- TVKCCA.** Three-view Kernel Canonical Correlation Analysis [59] (TVCCA for short) contains three views: two modalities views and a semantic views. It aims to learn a common subspace for visual, textual and semantic information via kernel CCA to achieve a better separation of data that belong to different categories.
- SM.** Semantic Matching [5] (SM for short) is to map image and text to a high-level semantic subspace by using multi-class logistic regression. In this common semantic space, the natural correspondences between cross-modal representation can be captured.
- Deep-SM.** Deep Semantic Matching [46] (Deep-SM for short) is a deep learning based method which maps image and text to a common semantic space by using CNN and LDA.

- **RE-DNN.** Regularized Deep neural network [11] (RE-DNN for short) to generate high-level representations of different modalities. This approach learns a joint model that captures both intra-modal and inter-modal relationships by an intra-modal regularization.
- **Corr-AE.** Correspondence Autoencoder [10] (Corr-AE for short) is built by correlating hidden representations with only common information of two uni-modal autoencoder. It incorporates representation learning and correlation learning into a single process.
- **ACMR.** Adversarial Cross-Modal Retrieval [15] (ACMR for short) is an adversarial learning based method to generate modality-invariant representations. This method minimizes the gap between different representations belong to the same categories, while maximizes the distances among semantically different data.

**Performance Metrics.** In our experiments, two cross-modal retrieval tasks are considered: image-to-text (I2T) retrieval and text-to-image (T2I) retrieval. The I2T retrieval is to retrieve similar texts from an image query, and the T2I retrieval is to find similar images from a text query. To evaluate cross-modal correlation learning, Pearson correlation coefficient is used to measure the correlation between different representations:

$$\begin{aligned} \text{Pearson}(\mathbf{X}, \mathbf{Y}) &= \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\sigma_X \sigma_Y} \\ &= \frac{\sum_{i=1}^n (X_i - \mu_X) \times (Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (X_i - \mu_X)^2} \times \sqrt{\sum_{i=1}^n (Y_i - \mu_Y)^2}} \end{aligned} \quad (35)$$

To comprehensive evaluate the retrieval performance, Precision-Recall curves (PR-Curves) are used in these experiments, in which

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

Besides, based on precision and recall, mean Average Precision scores (mAP) are calculated for the global performance evaluation. mAP is an important measurement to evaluate the global performance of retrieval algorithm, which is calculated via the following equation:

$$\begin{aligned} mAP &= \frac{1}{|Q|} \sum_{Q \in Q} AP(Q), \\ AP(Q) &= \frac{1}{N} \sum_{r=1}^R P_Q(r) \Phi(r), \end{aligned} \quad (36)$$

where  $Q$  is the query set,  $N$  is the number of results that is correlative to the query,  $R$  is number of results,  $P_Q(r)$  is the precision of the current query  $Q$ ,  $\Phi(r)$  is an indicator function. If the  $r$ -th result is correlative to the query,  $\Phi(r) = 1$ , otherwise  $\Phi(r) = 0$ .

**Experimental Environment.** All the experiments are run on a workstation with Intel(R) CPU Xeon 2.60GHz, 16GB RAM and NVIDIA GeForce GTX 1080 GPU with 8GB memory running Ubuntu 16.04 LTS Operation System. All algorithms in the experiments are implemented in Python.

## 5.2. Implementation Details

In our experiments, VisNet is realized by a pre-trained AlexNet whose the model parameters are provided by [60]. Then, we fine-tune this model on Wikipedia, NUS-WIDE and Pascal Sentence datasets. Specifically, the first five convolutional layers are initialized by the pre-training parameters, and the last two fully-connected layers are initialized by a Gaussian distribution  $G(\mu, \sigma^2)$ , where  $\mu = 0, \sigma = 0.01$ . All the training samples are resized to  $256 \times 256 \times 3$  without cropping. For the training, different learning rates are set to different layers: for the first five convolutional layer, the learning rate is set to 0.001; for fully-connected layers, the learning rate is set to 0.01. stochastic



Methods	Wikipedia			NUS-WIDE			Pascal Sentence		
	Img2Txt	Txt2Img	Average	Img2Txt	Txt2Img	Average	Img2Txt	Txt2Img	Average
CCA	21.01	17.84	19.43	38.17	36.80	37.49	11.21	12.06	11.64
DCCA	29.58	28.12	28.85	41.50	36.82	39.16	15.83	15.10	15.47
TVKCCA	20.13	22.06	21.09	39.50	40.39	39.95	14.85	15.42	14.89
SM	23.34	28.51	25.93	39.16	42.37	40.77	18.74	21.12	20.14
Deep-SM	39.90	35.43	37.67	57.80	62.55	60.18	44.63	48.05	46.34
RE-DNN	28.23	24.25	26.24	39.25	41.82	40.54	20.28	22.54	21.41
Corr-AE	32.63	36.10	34.37	31.93	38.63	35.28	29.86	28.35	29.36
ACMR	50.33	61.71	56.02	58.41	57.85	58.13	60.48	59.75	60.12
ALSCOR	51.05	63.58	57.32	60.88	65.26	63.07	63.92	68.41	66.17

**Table 2.** The performance (mAP @50 IN %) of the proposed method and the compared methods on Wikipedia, NUS-WIDE and Pascal Sentence for image-to-text (Img2Txt) retrieval, text-to-image (Txt2Img) retrieval and average performance.

gradient descent (SGD) algorithm with a mini-batch size of 128 is used to optimize the training, and the momentum is set to 0.8, the weight decay is set to 0.0005 to reduce the training error, the dropout ratio in each layer is set to 0.5. For the training of TxtNet, word2vec model is implemented by Skip-gram. We pre-train this model on Wikipedia corpus by using SGD algorithm, the margin is set to 0.2. The word2vec model outputs 100-dimensional word embeddings from the input texts, which are fed into the BiLSTM network. Following the settings in [61], the dimension of LSTM output vectors is 141 for one direction. On all datasets, the mini-batch size is set to 100.

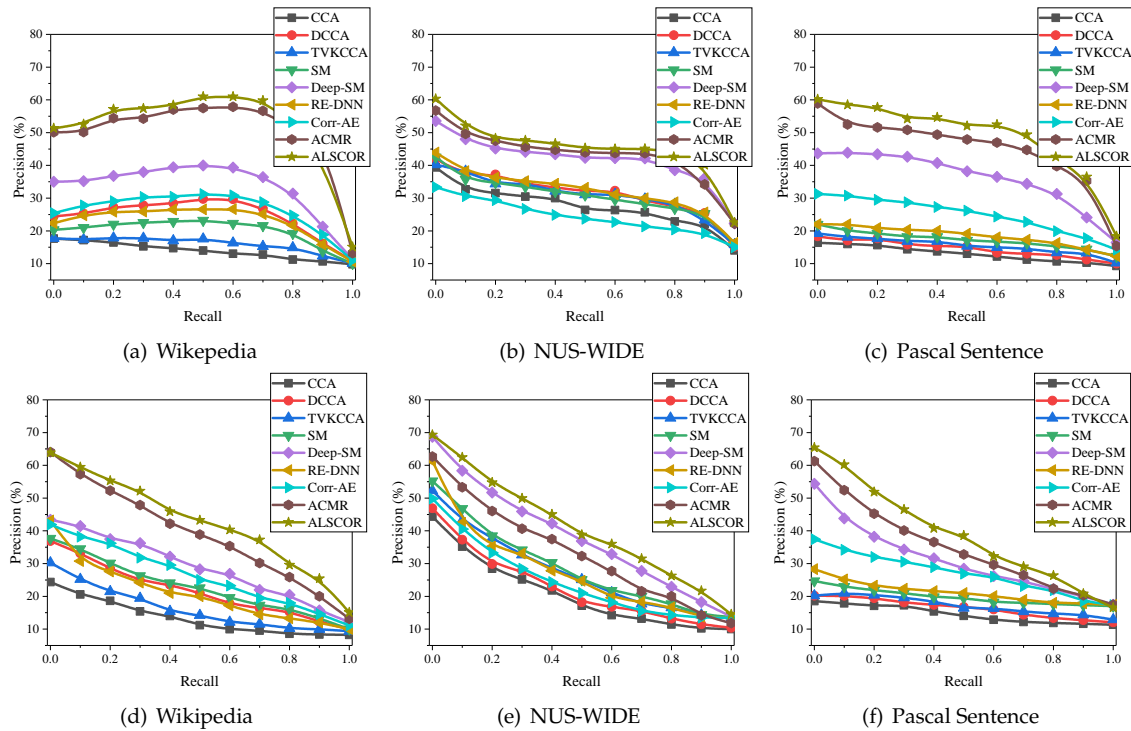
### 5.3. Performance Evaluation

In this section, we show the performance evaluation of the proposed method on Wikipedia, NUS-WIDE, and Pascal Sentence datasets, and compared it with 8 cross-modal retrieval approaches. The experimental results are reported in Table 2 and Fig. 5.

**Experiments on Wikipedia Dataset.** We compare the proposed ALSCOR with CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, ACMR on Wikipedia dataset. From Table 2 we can see that for both Img2Txt and Txt2Img task, our method (Img2Txt=51.05%, Txt2Img=63.58%) not only defeats the traditional approaches such as CCA (Img2Txt=21.01%, Txt2Img=17.84%), DCCA (Img2Txt=29.58%, Txt2Img=28.12%), TVKCCA (Img2Txt=20.13%, Txt2Img=22.06%), SM (Img2Txt=23.34%, Txt2Img=28.51%), etc., but also outperform the adversarial learning based method ACMR (Img2Txt=50.33%, Txt2Img=61.71%). This is mainly because the proposed ALSCOR has more powerful semantic representation. That means our method can capture more abstract concepts information to bridge the semantic gap. On the other hand, the performance of adversarial learning based methods, i.e., ACMR and ALSCOR, are much better than others, which verifies the adversarial learning can strongly support the common semantic subspace learning.

Fig. 5 (a) and Fig. 5 (d) illustrate the Precision-Recall curves of CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, ACMR and ALSCOR on Wikipedia dataset. From Fig. 5 (a), it is obvious that with the increasing of recall, the precision of ALSCOR and ACMR rise gradually and then drop rapidly near recall=1.0. Unsurprisingly, they perform much better than others over all values of recall. This verifies the improvement brought from adversarial learning once again. Similar to the discussed above, the precision of ALSCOR is a bit higher than ACMR for both Img2Txt retrieval and Txt2Img retrieval, which is because the advanced cross-modal representation learning model (the combination of VisNet and TxtNet).

**Experiments on NUS-WIDE Dataset.** We evaluate the performance of CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, ACMR and the proposed ALSCOR on NUS-WIDE dataset, shown in the middle column of Table 2. Similar to the experiment on Wikipedia, our method outperforms all the opponents by 60.88% for Img2Txt retrieval and 65.26% for Txt2Img retrieval. ACMR is the second best method, whose performance (Img2Txt=58.41%, Txt2Img=57.85%) is close to our method.



**Figure 5.** Precision-Recall curves for image-to-text query and text-to-image query on Wikipedia, NUS-WIDE and Pascal Sentence dataset. (a) Wikipedia Img2Txt Query. (b) NUS-WIDE Img2Txt Query. (c) Pascal Sentence Img2Txt Query. (d) Wikipedia Txt2Img Query. (e) NUS-WIDE Txt2Img Query. (f) Pascal Sentence Txt2Img Query.

Different from the experiment on Wikipedia, Deep-SM performs better, which achieve Img2Txt=57.80% and Txt2Img=62.55%. However, it still cannot defeat ALSCOR. Meanwhile, the performance of other methods is far behind the proposed method.

Fig. 5 (b) and Fig. 5 (e) show the comparison of ALSCOR and other 8 approaches on NUS-WIDE dataset for Img2Txt and Txt2Img, respectively. Fig. 5 (b) tells us that for Img2Txt retrieval, the precision of ALSCOR, ACMR and Deep-SM are close, which are higher than other five approaches obviously. Specifically, in the recall internal  $[0.1, 0.8]$ , these top-3 approaches are not sensitive to the changing of recall. Like the situation on Wikipedia, the performance of ALSCOR is better than ACMR. On the other hand, for Txt2Img retrieval, the performance gap between the top-3 methods and others is not so obvious, and the precision of Deep-SM is a little bit higher than ACMR. However, ALSCOR is still the best for all value of recall. This verifies effectiveness of the proposed ALSCOR.

**Experiments on Pascal Sentence Dataset.** The experimental results of the proposed ALSCOR and CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, ACMR on Pascal Sentence dataset are shown in the right column of Table 2. On this multimedia dataset, our method ALSCOR is still the best. It achieves  $mAP = 63.92\%$  for Img2Txt task and  $mAP = 68.41\%$  for Txt2Img task, which are obvious higher than ACMR ( $mAP = 63.92\%$  for Img2Txt and  $mAP = 68.41\%$  for Txt2Img), the most competitive opponent. The precision of other hand-crafted feature based methods, i.e., CCA, TVKCCA and SM, etc., are much lower than the two former. On the other hand, similar to the above experiments, the precision of ALSCOR for Txt2Img is better than Img2Txt. This is mainly because the textual semantics is easier to be learnt than images, and the combination of BiLSTM and CNN model can capture more high-level concept information from texts.

Fig. 5 (c) and Fig. 5 (f) demonstrates the trend of precision of ALSCOR and the compared methods with the varying of recall on Pascal Sentence dataset. For Img2Txt retrieval, we can see from Fig. 5 (c) that the proposed the precision of ALSCOR declines step-by-step with the rising of recall, which is higher than ACMR. Similar to the experiments on Wikipedia and NUS-WIDE, these two adversarial

learning based approaches are much more powerful than others for Img2Txt retrieval. On the other hand, in Fig. 5 (f), the superiority of them is not so obvious for Txt2Img retrieval, but they are still defeat other methods. Compared with ACMR, the proposed ALSCOR performs better.

## 6. Conclusions

In this paper, a novel cross-modal approach, adversarial learning based semantic correlation representation (ALSCOR) is proposed to address cross-modal retrieval problem. This approach is a combination of adversarial learning and cross-modal correlation learning. To bridge the semantic gap between different modalities, a deep learning based cross-modal correlation learning model is developed, which is integrated two branches (VisNet and TxtNet) to learn cross-modal representations and uses CCA model to learn the cross-modal correlation. Besides, a modality classifier is utilized to implement adversarial learning, which is to learn modality-invariant representations. In addition, an intra-modal classifier is used to capture the intra-modal discriminant information. We conduct comprehensive experiments on three benchmark datasets to evaluate the performance of the proposed method and compare it with 8 state-of-the-arts. Experimental results shows that the proposed ALSCOR has better performance than the state-of-the-arts.

**Author Contributions:** Conceptualization, L.Z. and J.L.; methodology, L.Z.; software, J.S. and X.W.; validation, L.Z. and X.W.; formal analysis, L.Z. and J.L.; investigation, J.S.; data curation, L.Z.; writing—original draft preparation, L.Z.; writing—review and editing, L.Z. and J.L.; visualization, X.W.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L.

**Acknowledgments:** This work was supported in part by the National Natural Science Foundation of China (61702560, 61472450, 61972203), the Key Research Program of Hunan Province (2016JC2018), project (2018JJ3691) of Science and Technology Plan of Hunan Province, and the Research and Innovation Project of Central South University Graduate Students (2018zzts177).

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

## References

1. Wang, K., Yin, Q., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215.
2. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 423-443.
3. Liu, X., Hu, Z., Ling, H., & Cheung, Y. M. (2019). MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
4. Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation (No. 47). Sage.
5. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010, October). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 251-260). ACM.
6. Wang, S., Gu, X., Lu, J., Yang, J. Y., Wang, R., & Yang, J. (2014, August). Unsupervised discriminant canonical correlation analysis for feature fusion. In *2014 22nd International Conference on Pattern Recognition* (pp. 1550-1555). IEEE.
7. Zu, C., & Zhang, D. (2016). Canonical sparse cross-view correlation analysis. *Neurocomputing*, 191, 263-272.
8. Gong, Y., Ke, Q., Isard, M., & Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2), 210-233.
9. Zhang, B., Hao, J., Ma, G., Yue, J., Zhang, J., & Shi, Z. (2015). Mixture of probabilistic canonical correlation analysis. *Journal of Computer Research and Development*, 52(07), 1463-1476.
10. Feng, F., Wang, X., & Li, R. (2014, November). Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 7-16). ACM.
11. Wang, C., Yang, H., & Meinel, C. (2015, November). Deep semantic mapping for cross-modal retrieval. In *2015 IEEE 27th International conference on tools with artificial intelligence (ICTAI)* (pp. 234-241). IEEE.
12. Srivastava, N., & Salakhutdinov, R. (2012, July). Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop* (Vol. 79).

13. Cai, G., Feng, Y., & Lin, Q. (2017, December). Cross-modal retrieval based on deep correlated network. In 2017 3rd IEEE International Conference on Computer and Communications (ICCC) (pp. 1226-1231). IEEE.
14. Peng, Y., Huang, X., & Qi, J. (2016, July). Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In IJCAI (pp. 3846-3853).
15. Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017, October). Adversarial cross-modal retrieval. In Proceedings of the 25th ACM international conference on Multimedia (pp. 154-162). ACM.
16. Zhang, X., Lai, H., & Feng, J. (2018). Attention-Aware Deep Adversarial Hashing for Cross-Modal Retrieval. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 591-606).
17. Wen, X., Han, Z., Yin, X., & Liu, Y. S. (2019). Adversarial Cross-Modal Retrieval via Learning and Transferring Single-Modal Similarities. arXiv preprint arXiv:1904.08042.
18. Shang, F., Zhang, H., Zhu, L., & Sun, J. (2019). Adversarial cross-modal retrieval based on dictionary learning. *Neurocomputing*, 355, 93-104.
19. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689-696).
20. Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222-2230).
21. Mroueh, Y., Marcheret, E., & Goel, V. (2015, April). Deep multimodal learning for audio-visual speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2130-2134). IEEE.
22. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.
23. Wang, Y., Wu, L., Lin, X., & Gao, J. (2018). Multiview spectral clustering via structured low-rank matrix factorization. *IEEE transactions on neural networks and learning systems*, 29(10), 4833-4843.
24. Zhu, L., Long, J., Zhang, C., Yu, W., Yuan, X., & Sun, L. (2019). An Efficient Approach for Geo-Multimedia Cross-Modal Retrieval. *IEEE Access*, 7, 180571-180589.
25. Cukurova, M. (2019). Learning analytics as AI extenders in education: Multimodal machine learning versus multimodal learning analytics. In *Proceedings of the Artificial Intelligence and Adaptive Education Conference* (pp. 1-3).
26. Zhang, C., Lin, Y., Zhu, L., Zhang, Z., Tang, Y., & Huang, F. (2019). Efficient region of visual interests search for geo-multimedia data. *Multimedia Tools and Applications*, 78(21), 30839-30863.
27. Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics* (pp. 162-190). Springer, New York, NY.
28. Shao, J., Wang, L., Zhao, Z., & Cai, A. (2016). Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. *Neurocomputing*, 214, 618-628.
29. Wang, S., Zhuang, F., Jiang, S., Huang, Q., & Tian, Q. (2015). Cluster-sensitive Structured Correlation Analysis for Web cross-modal retrieval. *Neurocomputing*, 168, 747-760.
30. Jin, C., Mao, W., Zhang, R., Zhang, Y., & Xue, X. (2015, February). Cross-modal image clustering via canonical correlation analysis. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
31. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
32. Yakhnenko, O., & Honavar, V. (2009, April). Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (pp. 283-293). Society for Industrial and Applied Mathematics.
33. Putthividhy, D., Attias, H. T., & Nagarajan, S. S. (2010, June). Topic regression multi-modal latent dirichlet allocation for image annotation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3408-3415). IEEE.
34. Blei, D. M., & Jordan, M. I. (2003, July). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 127-134). ACM.
35. Jia, Y., Salzmann, M., & Darrell, T. (2011, November). Learning cross-modality similarity for multinomial data. In *2011 International Conference on Computer Vision* (pp. 2407-2414). IEEE.

36. Lu, X., Wu, F., Tang, S., Zhang, Z., He, X., & Zhuang, Y. (2013, July). A low rank structural large margin method for cross-modal ranking. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 433-442). ACM.
37. Yu, J., Cong, Y., Qin, Z., & Wan, T. (2012, November). Cross-modal topic correlations for multimedia retrieval. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 246-249). IEEE.
38. Zoghbi, S., Heyman, G., Gomez, J. C., & Moens, M. F. (2015). Cross-modal attribute recognition in fashion. In Proceedings of NIPS Multimodal Machine Learning Workshop.
39. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.
40. Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). MIT press.
41. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
42. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
43. Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., & Masquelier, T. (2018). STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99, 56-67.
44. Zhang, C., Lin, Y., Zhu, L., Liu, A., Zhang, Z., & Huang, F. (2019). CNN-VWII: An efficient approach for large-scale video retrieval by image queries. *Pattern Recognition Letters*, 123, 82-88.
45. Wu, L., Wang, Y., & Shao, L. (2018). Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4), 1602-1612.
46. Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., & Yan, S. (2016). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2), 449-460.
47. He, Y., Xiang, S., Kang, C., Wang, J., & Pan, C. (2016). Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7), 1363-1377.
48. Shen, Y., Liu, L., Shao, L., & Song, J. (2017). Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4097-4106).
49. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., & Gao, X. (2017, February). Pairwise relationship guided deep hashing for cross-modal retrieval. In Thirty-First AAAI Conference on Artificial Intelligence.
50. Cao, Y., Long, M., Wang, J., & Liu, S. (2017, February). Collective deep quantization for efficient cross-modal retrieval. In Thirty-First AAAI Conference on Artificial Intelligence.
51. Hu, D., Wang, C., Nie, F., & Li, X. (2019, May). Dense Multimodal Fusion for Hierarchically Joint Representation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3941-3945). IEEE.
52. Gu, J., Cai, J., Joty, S. R., Niu, L., & Wang, G. (2018). Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7181-7189).
53. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
54. Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013, February). Deep canonical correlation analysis. In International conference on machine learning (pp. 1247-1255).
55. Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009, July). NUS-WIDE: a real-world web image database from National University of Singapore. In Proceedings of the ACM international conference on image and video retrieval (p. 48). ACM.
56. Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010, June). Collecting image annotations using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (pp. 139-147). Association for Computational Linguistics.
57. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
58. Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In null (p. 1470). IEEE.
59. Gong, Y., Ke, Q., Isard, M., & Lazebnik, S. (2012). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2), 210-233.



60. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678).
61. Tan, M., Santos, C. D., Xiang, B., & Zhou, B. (2015). Lstm-based deep learning models for non-factoid answer selection. arXiv preprint arXiv:1511.04108.