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

Lei Zhu  
Central South University  

Jiayu Song  
Central South University  

Xiaofeng Zhu  
University of Electronic Science and Technology of China  

Chengyuan Zhang  
Hunan University  

Shichao Zhang  
Central South University  

Xinpan Yuan  
Hunan University of Technology  

Abstract—Cross-modal retrieval has become a hot issue in past years. Many existing works pay attentions on correlation learning to generate a common subspace for cross-modal correlation measurement, and others use adversarial learning technique to abate the heterogeneity of multimodal data. However, very few works combine correlation learning and adversarial learning to bridge the intermodal semantic gap and diminish cross-modal heterogeneity. This article proposes a novel cross-modal retrieval method, named Adversarial Learning based Semantic COrrelation Representation (ALSCOR), which is an end-to-end framework to integrate cross-modal representation learning, correlation learning, and adversarial. Canonical correlation analysis model, combined
with VisNet and TxtNet, is proposed to capture cross-modal nonlinear correlation. Besides, intramodal classifier and modality classifier are used to learn intramodal discrimination and minimize the intermodal 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.

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, audios, etc., are generated, collected, stored, and shared on the Internet and online social networks, as shown in Figure 1. These multimodal data are usually used to describe the same events, scenes, or objects in our daily life, and users always have the requirements to search multimedia data by multimodal queries. This retrieval paradigm is called cross-modal retrieval, which attracts more and more attentions in the community of multimedia.\(^\text{18}\)

In the past decade, lots of approaches have been proposed to address cross-modal retrieval problem. The main challenge 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)\(^\text{9}\) is a widely used statistical method to find the correlations between cross-modal representations. Following the work by Rasiwasia \textit{et al.}\(^\text{1}\) several CCA-based research works have been presented to support cross-modal retrieval. Inspired by deep learning, a number of researchers exploit DNN to improve the performance of retrieval by learning the nonlinear correlations between modalities.

**Motivation.** The previous approaches aim to learn a common cross-modal semantic subspace. However, very few works combine correlation learning and adversarial learning to bridge the intermodal semantic gap and diminish cross-modal heterogeneity together. To overcome this challenge, for the first time, this article develops an end-to-end framework to bridge the semantic gap and diminish the cross-modal heterogeneity. Different from the existing studies,\(^\text{5,6}\) 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, termed as Adversarial Learning-based Semantic COrrelation Representation (ALSCOR). Inspired by deep CCA, we design a cross-modal deep representation CCA model that consists of a two-branch network. The CCA model accompanied by this two-branch model is used to learn the intermodal correlation. Besides, an intramodal classifier is used to learn the intramodal discriminative information. In addition, 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 article can be summarized as follows.

- We propose a novel framework that is a combination of cross-modal correlation learning and adversarial learning.
- To learn the intermodal nonlinear correlation, a two-branch cross-modal correlation model is developed, which is an integration of VisNet, TxtNet, and CCA. An intramodal classifier is utilized to learn the intramodal 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 eight state-of-the-arts. The results demonstrate that our method has great performance for cross-modal retrieval.

**RELATED WORK**

**Cross-Modal Retrieval**

Cross-modal retrieval is a significant problem in the area of multimedia computing,\(^\text{7,8}\) 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\(^9\) 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, the work by Rasiwasia et al.\(^1\) is the first work using CCA to address cross-modal retrieval. Wang et al.\(^2\) proposed a method called unsupervised discriminant CCA. Different from the existing research works, we combine deep CCA and adversarial learning method to learn better representations with modality invariance, which can implement feature distribution alignment between different modalities.

Deep Learning

As a very powerful technique, deep learning\(^10–12\) is used to overcome several challenges of multimedia analysis and retrieval, such as image classification, object recognition, video retrieval, multimodal/cross-modal retrieval, etc. Wei et al.\(^13\) proposed to use deep CNN to learn visual representations for cross-modal retrieval, which performs much better than traditional hand-crafted features.

Inspired by Goodfellow et al.\(^14\) several studies utilized adversarial learning to improve the cross-modal representation learning. Wang et al.\(^5\) are the first to use adversarial learning method to address cross-modal retrieval problem. Wen et al.\(^6\) introduced a new cross-modal similarity transferring method by adversarial learning. 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.

PRELIMINARY

In this section, we formalize the definition of cross-modal retrieval and the related notions. Besides, the cross-modal correlation measurement is proposed.

Problem Definition

**Definition 1 (Cross-Modal Retrieval).** Without losing generality, consider a multimedia dataset that contains multimodal data is denoted as \(D = \{X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n\}\), where \(X\) and \(Y\) denote two different modalities. Cross-modal retrieval aims to return a set of data \(R\) of one modality, which are correlative enough to the query of another modality, namely,
\[ R = \{ Y | Y \in D, Y' \in D \setminus R \}, \]
\[ \text{Corr}(X_Q, Y) \geq \text{Corr}(X_Q, Y') \]  \quad (1)

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

This article 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 an 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

\[ R_{I \rightarrow T} = \{ T | T \in D, T' \in D \setminus R_{I \rightarrow T} \}, \]
\[ \text{Corr}(I_Q, T) \geq \text{Corr}(I_Q, T') \]  \quad (2)

\[ R_{T \rightarrow I} = \{ I | I \in D, I' \in D \setminus R_{T \rightarrow I} \}, \]
\[ \text{Corr}(T_Q, I) \geq \text{Corr}(T_Q, I') \]  \quad (3)

Suppose that the multimedia dataset \( D = \{ (I_1, T_1), (I_2, T_2), \ldots, (I_n, T_n) \} \) contains \( n \) pairs of image and text. Each pair has a classification label vector denoted as \( L_i = \{ L_i^{(1)}, L_i^{(2)}, \ldots, L_i^{(c)} \} \in \mathbb{R}^c \), 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 \( \xi_i = (\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(\gamma_i)}) \in \mathbb{R}^{\gamma_i} \) and \( \xi_i = (\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(\gamma_T)}) \in \mathbb{R}^{\gamma_T} \), where \( \gamma_i \) and \( \gamma_T \) are the number of dimensions, and generally \( \gamma_i \neq \gamma_T \). Therefore, for the multimedia dataset \( D \), the image representation matrix, the text representation matrix and the classification label matrix are denoted as \( M_I = (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^{\gamma_I \times n} \), \( M_T = (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^{\gamma_T \times n} \), and \( M_L = (L_1, L_2, \ldots, L_n) \in \mathbb{R}^{c \times n} \), 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 dimensions of feature representations are different, and they are hard to be represented in the same distribution. Besides, the semantic concepts of representations are hard to be aligned. These two limitations hinder the correlation measurement of cross-modal data. Thus, two cross-modal mappings, \( \Omega_I((I_1, I_2, \ldots, I_n)) \) and \( \Omega_T((T_1, T_2, \ldots, T_n)) \) need to be learnt to project images \( (I_1, I_2, \ldots, I_n) \) and texts \( (T_1, T_2, \ldots, T_n) \) 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,

\[ \Omega_I((I_1, I_2, \ldots, I_n)) : \]
\[ (I_1, I_2, \ldots, I_n) \rightarrow (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^{\gamma \times n} \]  \quad (4)

\[ \Omega_T((T_1, T_2, \ldots, T_n)) : \]
\[ (T_1, T_2, \ldots, T_n) \rightarrow (\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n) \in \mathbb{R}^{\gamma \times n} \]  \quad (5)

where \( (\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n) \) and \( (\xi_1, \xi_2, \ldots, \xi_n) \) are the representation matrix of images and texts in the common semantic subspace. Therefore, the correlations between multimodal 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 \( I \), and a text \( T \), the representations of \( I \) and \( T \) in the common semantic subspace are \( \xi_i \) and \( \xi_j \), the cross-modal correlation between \( I \) and \( T \) is measured by the following equation:

\[
\text{Corr}(I_i, T_j) = \frac{\sum_{k=1}^{\gamma} (\xi_i^{(k)} - \mu_{\xi_i}) (\xi_j^{(k)} - \mu_{\xi_j})}{\sqrt{\sum_{k=1}^{\gamma} (\xi_i^{(k)} - \mu_{\xi_i})^2} \sqrt{\sum_{k=1}^{\gamma} (\xi_j^{(k)} - \mu_{\xi_j})^2}}
\]  \quad (6)

where \( \mu_{\xi_i} \) and \( \mu_{\xi_j} \) are the averages of \( \xi_i \) and \( \xi_j \), respectively.

**METHOD**

To learn a common subspace and bridge the semantic gap between different modalities, we propose an effective end-to-end framework, named ALSCOR, which is a combination of cross-modal correlation learning and adversarial learning.  

**Architecture of ALSCOR**

**Overview.** The general architecture of the proposed ALSCOR is illustrated in Figure 2, 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 an 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 pairs 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 deep representations of images. The other is TtxtNet, which is generated by text representations via a combination 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. Compared with deep fully connected neural networks in the deep CCA method, CNN is more powerful to capture high-level visual semantic information from images. In our approach, 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 layer 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 $y_{f} = 4096$.

**VisNet.** VisNet is a CNN-based model to generate deep representations of images, formally, $(\xi^{(1)}_i, \xi^{(2)}_i, \ldots, \xi^{(g)}_i) = \text{VisNet}(I_i; \theta_i)$, where $\theta_i$ is the model parameter vector. Compared with deep fully connected neural networks in the deep CCA method, CNN is more powerful to capture high-level visual semantic information from images. In our approach, 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 layer 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 $y_{f} = 4096$.

**TtxtNet.** TtxtNet is a combination of word2vec, BiLSTM network, and CNN model, which is to generate $\lambda$-dimensional text representations, namely $(\xi^{(1)}_i, \xi^{(2)}_i, \ldots, \xi^{(g)}_i) = \text{TtxtNet}(T_i; \theta_T)$, where $\theta_T$ is the model parameter vector. The word2vec model receives the inputs and generates a set of word embeddings $(v_1, v_2, \ldots, v_n)$, where $n$ is the number of the words. Compared to the traditional representation method (such as BoW),
word2vec can capture both semantic and syntactic 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

\[
\begin{align*}
i(t) &= \sigma(W_i[h(t-1), v(t)] + b_i) \\
 f(t) &= \sigma(W_f[h(t-1), v(t)] + b_f) \\
o(t) &= \sigma(W_o[h(t-1), v(t)] + b_o) \\
 \tilde{C}(t) &= \tanh(W_o[h(t-1), v(t)] + b_o) \\
 C(t) &= i(t) \times \tilde{C}(t) + f(t) \times C(t-1)
\end{align*}
\]

where \( h(t) \) is the hidden vector, \( \sigma \) is the sigmoid function, \( W_i, W_f, W_o, b_i, b_f, \) and \( b_o \) 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., \( K = \{ K^{(1)}, K^{(2)}, \ldots, K^{(\kappa)} \} \). For the \( j \)th location of the input map, the calculation can be formalized as follows:

\[
f^{(j)}_i(t) = g(\sum_{i=1}^{m} h(t+i-1)^T \ast K^{(i)} + b)
\]

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

\[
f_i(t) = (f^{(1)}_i(t), f^{(2)}_i(t), \ldots, f^{(H-m+1)}_i(t))
\]

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

\[
(f_1(t), f_2(t), \ldots, f_m(t)) = \max(f_1(t), f_2(t), \ldots, f_m(t))
\]

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

To restrain 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:

\[
(\xi^{(1)}_i, \xi^{(2)}_i, \ldots, \xi^{(\kappa)}_i) = W_{fc} \times (f_1(t), f_2(t), \ldots, f_m(t))^T \odot \mathcal{M} + \bm{\beta}
\]

where \( W_{fc} \) and \( \bm{\beta} \) are the parameters of the fully connected layers, \( \odot \) is the elementwise multiplication operator, and \( \mathcal{M} \) is a masking vector of Bernoulli random variables.

### 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 intramodal classifier is to learn the intramodal 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 an end-to-end nonlinear correlation learning model to maximize the cross-modal correlation. According to deep CCA, the correlation loss is formalized as follows:

\[
L_{corr}(I_i, T_i; \theta_I, \theta_T) = \text{Corr}(\text{VisNet}(I_i; \theta_I), \text{TxtNet}(T_i; \theta_T)).
\]

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

\[
(\theta^*_I, \theta^*_T) = \underset{(\theta_I, \theta_T)}{\arg \max} L_{corr}(I_i, T_i; \theta_I, \theta_T) \quad \forall (I_i, T_i) \in \mathcal{D}.
\]

The optimal parameters \( (\theta^*_I, \theta^*_T) \) are calculated by using gradient of the correlation objective on the training set \( \mathcal{D} \).
Discrimination Loss. The intramodal classifier is to maintain the discrimination of multimodal data after the cross-modal nonlinear 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 receive the representations of different modalities 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:

\[
L_{\text{disc}}(I_i; T_i; \theta_I, \theta_T, \theta_D) = -\frac{1}{m} \sum_{i=1}^{m} \log P(\text{VisNet}(I_i; \theta_I)) + \log P(\text{TxtNet}(T_i; \theta_T))
\]  

(19)

where \(P(\cdot)\) is the probability distribution of categories, \(\theta_D\) is the parameter vector of the classifier, and \(m\) is the number of samples in each minibatch 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 Wang et al.\(^5\), 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:

\[
L_{\text{adv}}(I_i; T_i; \theta_I, \theta_T, \theta_A) = -\frac{1}{n} \sum_{i=1}^{n} m_i \log D(\text{VisNet}(I_i; \theta_I); \theta_A) + \log (1 - D(\text{TxtNet}(T_i; \theta_T); \theta_A))
\]  

(20)

where \(m_i\) is the ground-truth modality label of each representation.

Adversarial Learning. For the adversarial learning process, we incorporate the correlation loss (17), discrimination loss (19), and adversarial loss (20), and optimize them as a min–max game, shown as follows:

\[
\begin{align*}
\theta_I^* &= \arg \min_{\theta_I} (\alpha L_{\text{corr}}(I_i; T_i; \theta_I, \theta_T) + \delta L_{\text{disc}}(I_i; T_i; \theta_I, \theta_T, \theta_D)) \\
&\quad + \epsilon L_{\text{adv}}(I_i; T_i; \theta_I, \theta_T, \theta_A)), \\
\theta_T^* &= \arg \max_{\theta_T} (\alpha L_{\text{corr}}(I_i; T_i; \theta_I, \theta_T) + \delta L_{\text{disc}}(I_i; T_i; \theta_I, \theta_T, \theta_D)) \\
&\quad - \epsilon L_{\text{adv}}(I_i; T_i; \theta_I, \theta_T, \theta_A)), \\
\theta_A^* &= \arg \max_{\theta_A} (\alpha L_{\text{corr}}(I_i; T_i; \theta_I, \theta_T) + \delta L_{\text{disc}}(I_i; T_i; \theta_I, \theta_T, \theta_D)) \\
&\quad - \epsilon L_{\text{adv}}(I_i; T_i; \theta_I, \theta_T, \theta_A))
\end{align*}
\]  

(21)

where \(\alpha, \delta, \text{and } \epsilon\) are the weight parameters for these three loss terms. The training can be realized by using a stochastic gradient descent algorithm.

EXPERIMENTS

In this section, we present the performance evaluation of the proposed method and the comparison with several state-of-the-arts on three multimedia datasets.

Experimental Setup

Datasets. Our experiments are conducted on three benchmark multimedia datasets: Wikipedia, NUS-WIDE, and Pascal Sentence, which are widely used in multimodal/cross-modal retrieval tasks. Some samples of these datasets are shown in Figure 3.

Compared Approaches. To verify the effectiveness of the proposed method, we compare the proposed ALSCOR with eight approaches, i.e., CCA\(^9\), DCCA\(^15\), TVKCCA\(^17\), SM\(^1\), Deep-SM\(^13\), REDNN\(^4\), Corr-AE\(^3\), and ACMR\(^5\).

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. To comprehensive evaluate the retrieval performance, precision–recall curves (PR curves) and mean Average Precision scores (mAP) are used in these experiments.

Implementation Details. For VisNet, all the images are resized to 256 \times 256 \times 3 without
cropping. 224 × 224 × 3 patches (and their horizontal reflections) are extracted randomly. We set different learning rates to different layers: the learning rates of first five convolutional layers and the FC-layer are set to 0.001 and 0.01, respectively. ForTxtNet, the word2vec model is pretrained by Skip-gram model, which generates 300-dimensional word vectors from texts. The filter size of CNN is set to 3 × 300, the dropout rate is set to 0.5.

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

**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 eight cross-modal retrieval approaches. The experimental results are reported in Table 1 and Figure 4.

**Experiments on Wikipedia Dataset.** We compare the proposed ALSCOR with CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, and ACMR on Wikipedia dataset. From Table 1, 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=25.93%), etc., but also outperforms 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.

Figure 4(a) and (d) illustrates the PR curves of CCA, DCCA, TVKCCA, SM, Deep-SM, RE-DNN, Corr-AE, ACMR, and ALSCOR on Wikipedia dataset. From Figure 4(a), it is obvious that with the increase in 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 earlier, the precision of ALSCOR is a bit higher than ACMR for both Img2Txt retrieval and Txt2Img retrieval, which is because of the advanced cross-modal representation learning model (the combination of VisNet andTxtNet).

**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. 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. On the other hand, the mAPs of all these methods on NUS-WIDE are higher than Wikipedia. That mainly because the categories in Wikipedia are more abstract, such as "art," "literature," "royalty," etc., which means the discriminant features are hardly to learn.

### Table 1. Comparison Results (mAP@50 in %) With Eight State-of-the-Arts on Wikipedia Dataset.

| Method | Img2Txt | Txt2Img | Average |
|--------|---------|---------|---------|
| CCA    | 21.01   | 17.84   | 19.43   |
| DCCA   | 29.58   | 28.12   | 28.85   |
| TVKCCA | 20.13   | 22.06   | 21.09   |
| SM     | 23.34   | 28.51   | 25.93   |
| Deep-SM| 39.90   | 35.43   | 37.67   |
| RE-DNN | 28.23   | 24.25   | 26.24   |
| Corr-AE| 32.63   | 36.10   | 34.37   |
| ACMR   | 50.33   | 61.71   | 56.02   |
| ALSCOR | **51.05** | **63.58** | **57.32** |

The best performance values are in bold-font.

### Table 2. Comparison Results (mAP@50 in %) With Eight State-of-the-Arts on NUS-WIDE Dataset.

| Method | Img2Txt | Txt2Img | Average |
|--------|---------|---------|---------|
| CCA    | 38.17   | 36.80   | 37.49   |
| DCCA   | 41.50   | 36.82   | 39.16   |
| TVKCCA | 39.50   | 40.39   | 39.95   |
| SM     | 39.16   | 42.37   | 40.77   |
| Deep-SM| 57.80   | 62.55   | 60.18   |
| RE-DNN | 39.25   | 41.82   | 40.54   |
| Corr-AE| 31.93   | 38.63   | 35.28   |
| ACMR   | 58.41   | 57.85   | 58.13   |
| ALSCOR | **60.88** | **65.26** | **63.07** |

The best performance values are in bold-font.
Figure 4(b) and (e) shows the comparison of ALSCOR and other eight approaches on NUS-WIDE dataset for Img2Txt and Txt2Img, respectively. Figure 4(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-three 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-three 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, and ACMR on Pascal Sentence dataset are shown in the right column of Table 3. 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 obviously 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, SM, etc., are much lower than the two former. On the other hand, similar to the previous 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.

Figure 4(c) and (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 Figure 4(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 Figure 4(f), the superiority of them is not so obvious for Txt2Img retrieval, but they still defeat other methods. Compared with ACMR, the proposed ALSCOR performs better.

CONCLUSION
In this article, a novel cross-modal approach ALSCOR is proposed. This approach is a combination of adversarial learning and cross-modal correlation learning. To bridge the semantic gap, a deep-learning-based cross-modal correlation learning model is developed. Besides, a modality classifier is utilized to implement adversarial learning, which is to learn modality invariant representations. In addition, an intramodal classifier is used to capture the intramodal discriminant information. We conduct comprehensive experiments on three benchmark datasets to evaluate the performance of ALSCOR and eight state-of-the-arts. Experimental results shows that ALSCOR has better performance than the state of the arts. In our future studies, we will consider both intermodal and intramodal semantic correlations, which can further reduce the cross-modal semantic gap.

ACKNOWLEDGMENTS
This work was supported in part by the National Natural Science Foundation of China under Grant 61702560, Grant 61472450, and Grant 61972203, and in part by the Science and Technology Plan of Hunan Province under Project 2018JJ3691.

| Method  | Img2Txt | Txt2Img | Average |
|---------|---------|---------|---------|
| CCA     | 11.21   | 12.06   | 11.64   |
| DCCA    | 15.83   | 15.10   | 15.47   |
| TVKCCA  | 14.85   | 15.42   | 14.89   |
| SM      | 18.74   | 21.12   | 20.14   |
| Deep-SM | 44.63   | 48.05   | 46.34   |
| RE-DNN  | 20.28   | 22.54   | 21.41   |
| Corr-AE | 29.86   | 28.35   | 29.36   |
| ACMR    | 60.48   | 59.75   | 60.12   |
| ALSCOR  | 63.92   | 68.41   | 66.17   |

The best performance values are in bold-font.
REFERENCES

1. N. Rasiwasia et al., “A new approach to cross-modal multimedia retrieval,” in Proc. 18th ACM Int. Conf. Multimedia, Oct. 2010, pp. 251–260.

2. S. Wang, X. Gu, J. Lu, J. Y. Yang, R. Wang, and J. Yang, “Unsupervised discriminant canonical correlation analysis for feature fusion,” in Proc. IEEE 22nd Int. Conf. Pattern Recognit., Aug. 2014, pp. 1550–1555.

3. F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in Proc. 22nd ACM Int. Conf. Multimedia, Nov. 2014, pp. 7–16.

4. C. Wang, H. Yang, and C. Meinel, “Deep semantic mapping for cross-modal retrieval,” in Proc. IEEE 27th Int. Conf. Tools Artif. Intell., Nov. 2015, pp. 234–241.

5. B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 154–162.

6. X. Wen, Z. Han, X. Yin, and Y. S. Liu, “Adversarial cross-modal retrieval via learning and transferring single-modal similarities,” in Proc. IEEE Int. Conf. Multimedia Expo., 2019, pp. 478–483.

7. Y. Wang, L. Wu, X. Lin, and J. Gao, “Multiview spectral clustering via structured low-rank matrix factorization,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

8. L. Zhu, J. Long, C. Zhang, W. Yu, X. Yuan, and L. Sun, “An efficient approach for geo-multimedia cross-modal retrieval,” IEEE Access, vol. 7, pp. 180571–180589, 2019.

9. H. Hotelling, “Relations between two sets of variates,” in Breakthroughs in Statistics. New York, NY, USA: Springer, 1992, pp. 162–190.

10. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, 2015, Art. no. 436.

11. L. Wu, Y. Wang, J. Gao, and X. Li, “Where-and-when to look: Deep Siamese attention networks for video-based person re-identification,” IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.

12. Z. Zhang, Z. Tang, Z. Zhang, Y. Wang, J. Qin, and M. Wang, “Fully-convolutional intensive feature flow neural network for text recognition,” 2019, arXiv:1912.06446.

13. Y. Wei et al., “Cross-modal retrieval with CNN visual features: A new baseline,” IEEE Trans. Cybern., vol. 47, no. 2, pp. 449–460, Feb. 2017.

14. I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

15. G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in Proc. Int. Conf. Mach. Learn., Feb. 2013, pp. 1247–1255.

16. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vision, vol. 60, no. 2, pp. 91–110, 2004.

17. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multiview embedding space for modeling internet images, tags, and their semantics,” Int. J. Comput. Vis., vol. 106, no. 2, pp. 210–233, 2012.

18. Y. Wang, “Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion,” 2020, arXiv:2006.08159.

Lei Zhu’s research interests include machine learning and its applications, such as pattern recognition, computer vision, and multimedia information processing. He is currently working toward the Ph.D. degree in computer science and technology with the School of Computer Science and Engineering, Central South University, Changsha, China. Contact him at leizhu@csu.edu.cn.

Jiayu Song was born in Inner Mongolia Autonomous region, China. His main research interests include deep learning and machine learning. He has been working toward the bachelor’s degree in data science and big data technology with Central South University, Changsha, China, since 2017. Contact him at jiayusong@cau.edu.cn.

Xiaofeng Zhu’s current interests include mining useful knowledge or information from big multimedia or medical imaging data. He received the Ph.D. degree in computer science from The University of Queensland, Brisbane, QLD, Australia. Contact him at xfzhu0011@hotmail.com.

Chengyuan Zhang is currently an associate professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His main research interests include information retrieval, and query processing on spatial data and multimedia data. He received the Ph.D. degree in computer science from the University of New South Wales, Sydney, NSW, Australia, in 2011 and 2015 separately. He is the first corresponding author of this article. Contact him at cyzhang@csu.edu.cn.
Shichao Zhang is currently a China National Distinguished Professor with the School of Computer Science and Technology, Central South University, Changsha, China. His research interests include information quality and pattern discovery. He received the Ph.D. degree in computer science from Deakin University, Melbourne, VIC, Australia. He is the second corresponding author of this article. Contact him at zhangsc@csu.edu.cn.

Xinpan Yuan is currently a lecturer with the School of Computer, Hunan University of Technology, Zhuzhou, China. His current research interests include natural language processing and information retrieval. He received the Ph.D. degree in information science and engineering from Central South University, Changsha, China, in 2012. Contact him at xpyuan@hut.edu.cn.

Contact department editor Yang Wang at yangwang@hfut.edu.cn.