Adversarial Attentive Multi-Modal Embedding Learning for Image-Text Matching

KAIMIN WEI\(^1\) AND ZHIBO ZHOU\(^2\)

\(^1\)College of Information Science and Technology, Jinan University, Guangzhou 510632, China
\(^2\)School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Corresponding author: Zhibo Zhou (zhouzhibo@buaa.edu.cn)

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010137005, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2017A030313334, and in part by the Science and Technology Program of Guangzhou of China Grant 201802010061.

ABSTRACT Matching the image and text with deep models has been extensively studied in recent years. Mining the correlation between image and text to learn effective multi-modal features is crucial for image-text matching. However, most existing approaches model the different types of correlation independently. In this work, we propose a novel model named Adversarial Attentive Multi-modal Embedding Learning (AAMEL) for image-text matching. It combines adversarial networks and attention mechanism to learn effective and robust multi-modal embeddings for better matching between the image and text. Adversarial learning is implemented as an interplay between two processes. First, two attention models are proposed to exploit two types of correlation between the image and text for multi-modal embedding learning and to confuse the other process. Then the discriminator tries to distinguish the two types of multi-modal embeddings learned by the two attention models, in which the two attention models are reinforced mutually. Through adversarial learning, it is expected that both the two embeddings from the attention models can well exploit the two types of correlation, and thus they can deceive the discriminator that they are generated from the other attention-based model. By integrating the attention mechanism and adversarial learning, the learned multi-modal embeddings are more effective for image and text matching. Extensive experiments have been conducted on the benchmark datasets of Flickr30K and MSCOCO to demonstrate the superiority of the proposed approaches over the state-of-the-art methods on image-text retrieval.

INDEX TERMS Adversarial learning, attention, multi-modal, embedding, image-text matching.

I. INTRODUCTION

With the massive explosion of social multimedia on the Internet, multi-modal data containing various media types (e.g., image, text, and video) has been increasingly popular in people’s daily life. The joint modeling of multi-modal data gives rise to a wide variety of applications. Image captioning \([1]–[6]\) aims to find or generate a natural sentence to describe an image. Image search \([7]–[11]\) is to retrieve related images given a language sentence or several keywords. A solution to these applications is to learn an effective matching model such that the semantically related image-text pairs can be assigned higher matching scores than the unrelated ones.

The task of image-text matching \([12]–[18]\) has attracted broad interests in both industry and academia. However, it is still challenging due to that the multiple modalities of image and text have complicated correlation between each other. Specifically, it can be observed that there exist two types of fine-grained correlation between image and text, i.e., regions-text correlation and words-image correlation, which can be also considered as the bi-directional correlation. Figure 1 shows an example of image-text pair in the dataset of Flickr30K. On one side, it can be seen that the text has close correlation with some specific regions in the image, which denotes the regions-text correlation. For example, in Figure 1, the visual regions of the little boy and the football are the more important areas to reflect the semantic information of the text description. On the other side, the image is also closely related to some specific words in the text, which denotes the words-image correlation. In Figure 1, “boy”, “blue”, “shirt”, and “ball” are more meaningful words to cover the content of the picture. If these two types of correlation...
are parsed accurately, the image and text can be modeled jointly to learn salient multi-modal features for more effective matching.

There are two main lines of researches for image-text matching. The first is Canonical Correlation Analysis (CCA) based methods [19]. CCA computes a linear mapping to maximize the correlation between the two projected views. DCCA [14], [20] extends CCA by employing the high level visual and textual features extracted with deep neural networks. Another strategy of image-text matching is to use a ranking loss to conduct metric learning [13], [15], [16], [21]. In [21], an image-text representation learning method is proposed to build a two-branch network with both ranking loss and structure-preserving loss. DAN [22] is proposed to build a dual attention reasoning model for image-text matching. However, these methods cannot effectively capture the bi-directional correlation between image and text.

Attention mechanism aims to dynamically extract the salient features from the sequential or spatial data. It has been proved to be effective in many tasks of multi-modal data, such as visual question answering [23]–[25], multimodal sentiment analysis [26]–[28] and multimodal representation learning [29]–[33]. As shown in Figure 1, there exist two types of fine-grained correlation between image and text. Inspired by the recent works on attention mechanism, the visual attention model can be employed to exploit the regions-text correlation to learn a joint embedding named RT for the image-text pair. Meanwhile, the semantic attention model is employed to exploit the words-image correlation to learn the other joint embedding named WI. Since the two types of joint embeddings usually have inconsistent distributions, combining them directly is not effective to bridge the gap for image-text matching learning. The existing attention-based image-text matching approaches typically focus on one type of correlation or treat each type of correlation independently. However, as shown in Figure 1, both of the two types of correlation reflect the matching information of the image-text pair, and they are also complementary to each other. Therefore, these approaches fail to fully preserve the underlying relation of the two types of correlation. Preserving this relation and the structure between items would require that the gap between the two embeddings of RT and WI of the same image-text pair is minimized, while the distances between dis-matched items of the same type of the joint embedding are maximized.

To address the problems existed in the previous attention-based image-text matching methods, we propose a new approach Adversarial Attentive Multi-modal Embedding Learning (AAMEL) which is built around the concept of adversarial learning [34]. The core of this approach is the interplay between attentive embedding learning and embedding discriminating. Two attention models are used to learn the multi-modal embeddings of RT and WI by exploiting the two types of correlation. They have the objective to confuse the discriminator that acts as an adversary. On the other side, a discriminator is built to distinguish the two types of joint embeddings RT and WI learned from the two attention models, which also steers the learning of the joint embeddings. Through the adversarial process, it is expected that both the embeddings of RT and WI can well exploit the regions-text and words-image correlation, and thus they can deceive the discriminator that they are generated from the other attention-based model. In this way, the adversarial learning reinforces the two attention models mutually to make the learned multi-modal embeddings more effective for image-text matching. Then, a pairwise margin ranking loss is proposed to capture the triplet constraints for matching score learning. The main contributions are summarized as follows:

- We investigate to exploit the bi-directional correlation between image and text for image-text matching learning. This process is more effective to learn the joint embeddings compared to those considering only one type of correlation or treating each type of correlation independently.
- We propose a novel approach, i.e., Adversarial Attentive Multi-modal Embedding Learning (AAMEL), for image-text matching. To the best of our knowledge, this is the first work to learn the attentive multi-modal embeddings with the adversarial principle for image-text matching.
- The proposed approach AAMEL is evaluated on two benchmark datasets. The experimental results demonstrate the superiority of AAMEL.

The rest of this paper is organized as follows. Section II summarizes the related works. Then section III describes AAMEL in details. Experimental results are presented in section IV. Finally, we conclude this work in section V.
II. RELATED WORK
Existing works of image-text matching can be roughly classified into CCA-based methods and ranking-based methods.

Canonical Correlation Analysis (CCA) [19] aims to compute a linear mapping to maximize the correlation between two projected views. KCCA [19] is an extension of CCA which reproduces kernel hilbert spaces with corresponding kernels to capture the nonlinear relation between multi-view data. Klein et al. [35] shows that properly normalized CCA can achieve satisfying performance by utilizing state-of-the-art visual and textual features. Due to the limitation of these shallow models, it is difficult for these methods to capture the highly non-linear cross-modal relation. With the development of deep learning technique, deep CCA [14], [20], [36] is proposed to apply CCA upon the high level visual and textual features extracted by deep neural networks. Deep CCA allows an end-to-end learning manner to propagate the gradient down in a deep learning framework. However, as described in [37], the generalized eigenvalue problem of CCA cannot be well solved by stochastic gradient descent due to the unstable covariance estimation in mini-batch.

Another strategy for image-text matching is to use a ranking loss to perform metric learning on image-text pairs [13], [21]. DeVISE [38] is proposed to learn a linear transformation to project images and labels into a joint visual-semantic embedding space. In [21], an image-text representation learning method is proposed to build a two-branch network with both ranking loss and structure-preserving loss. Ma et al. [13] propose the multi-modal convolutional neural networks (mCNNs) to match image and sentence at different semantic levels. Karpathy et al. [39] explicitly compute all pairwise distances to automatically calculate the alignments between visual regions and textual fragments. CMDN [40] is proposed to build multiple deep networks for cross-media shared representation learning. It integrates the intra-modal and inter-modal representations to learn the cross-modal correlation using hierarchical neural networks. CHAIN-VSE [16] builds a bidirectional retrieval framework which relies on character-level inception module for visual-semantic embeddings. Mithun et al. [41] build a two-stage approach for image-text retrieval by using a supervised ranking loss combined with weakly-supervised web images to learn multimodal representation. However, these methods cannot well capture the fine-grained correlation between image and text.

There are also some multi-modal matching methods utilizing attention mechanism [15], [22] or adversarial networks [42]–[47]. sm-LSTM [15] learns an attention-based selective LSTM for matching. Dual Attention Networks for image-text matching (m-DAN) [22] is proposed to learn two branches of representations for image and text separately. The attention networks in m-DAN are built in two branches for two modalities, which neglects the cross-modal correlation between image and text. Cm-gans [42] is proposed to seek a common subspace with adversarial learning to classify different uni-modalities for representation learning. SCAN [43] is proposed to build two kinds of stacked cross attention for correlation learning and then average their predicted similarity scores for matching. However, this shallow combination is not effective to find the internal connections between the two models. Compared with SCAN, our method exploits the correlation from a more macroscopic perspective. Furthermore, our two attention models are not treated separately but combined with adversarial learning to reinforce the two attention models mutually to learn the multi-modal embeddings for matching. Gu et al. [44] integrate the image and text adversarial generative models into image-text searching. DAML [45] is proposed to use a deep metric learning framework with adversarial regularization between two modalities for cross-modal retrieval. ACMR [46] is proposed for image-text retrieval by seeking an effective common subspace with adversarial learning to discriminate between different modalities. Different from these works, our adversarial learning framework is built to guide the two attention models to learn from each other for more effective multi-modal embedding learning.

III. ADVERSARIAL ATTENTIVE EMBEDDING LEARNING

A. PROBLEM STATEMENT
In this section, we give an overview of the proposed model. AAMEL. Specifically, a visual attention model is employed to exploit the correlation between visual regions and text to learn the regions-text (RT) embedding. Meanwhile, a semantic attention model is employed to exploit the correlation between textual words and image to learn the words-image (WI) embedding. Then, the two attention models are reinforced mutually through adversarial learning for more effective multi-modal embedding learning. That is, a discriminator is used to discriminate the embeddings of RT and WI learned from the two attention models, while the two attention models aim to confuse the discriminator. Finally, a pairwise margin ranking loss is proposed to learn the matching score. In this way, it can be ensured that the learned multi-modal embeddings of RT and WI can both capture the two types of correlation more effectively and are also discriminative among different image-text pairs.

B. LEARNING REGIONS-TEXT EMBEDDING
Usually, the text has close correlation with some specific regions in the corresponding image. Mining this type of correlation is useful to learn more effective multi-modal representation for image-text matching. In this section, a visual attention model is proposed to exploit the correlation between
visual regions and text to learn the multi-modal representation named RT embedding.

Given an image $V_i$, we first extract the features of region maps with pre-trained deep CNNs as $R_i \in \mathbb{R}^{d \times m \times m}$, where $m \times m$ is the number of regions and $d$ is the dimensionality of the feature vector of each region. For the corresponding text $T_i$, each word is embedded with pre-trained word embeddings and then fed into an LSTM, and the output of the last cell is used as the text representation $H_i \in \mathbb{R}^c$.

To fuse region features $R_i$ and text features $H_i$ for correlation learning, we first project them into a $c$-dimensional common space as follows:

\begin{align}
R_{i}^{(c)} &= \tanh(W_r R_i + b_r), \quad R_{i}^{(c)} \in \mathbb{R}^{c \times m \times m} \\
H_{i}^{(c)} &= \tanh(W_h H_i + b_h), \quad H_{i}^{(c)} \in \mathbb{R}^c
\end{align}

where $W_r \in \mathbb{R}^{c \times d}$ and $W_h \in \mathbb{R}^{c \times h}$ are the trainable parameters, and $b_r$ and $b_h$ are the bias parameters. Then $R_{i}^{(c)}$ and $H_{i}^{(c)}$ are fused by element-wise multiplication:

\begin{align}
F_{i}^{(c)} &= R_{i}^{(c)} \odot (H_{i}^{(c)}) \odot 1, \quad F_{i}^{(c)} \in \mathbb{R}^{c \times m \times m}
\end{align}

where $1 \in \mathbb{R}^{c \times m \times m}$ is used to broadcast the dimensionality of $H_{i}^{(c)}$ to $c \times m \times m$, which matches the spatial size of the image. The symbol $\odot$ denotes the element-wise multiplication. The attention map is then calculated by convolving the fused representation $F_{i}^{(c)}$ with a $1 \times 1$ kernel followed by a softmax function over $m \times m$ regions:

\begin{align}
\alpha_i &= \text{softmax}(\tanh(W_a \ast F_{i}^{(c)} + b_a)), \quad \alpha_i \in \mathbb{R}^{m \times m}
\end{align}

where the symbol $\ast$ denotes the convolutional operation, $W_a \in \mathbb{R}^{c \times 1 \times 1}$ and $b_a \in \mathbb{R}^c$ are the convolutional parameters. Then, the attentive features of the image can be calculated as the weighted average of the visual features over all regions:

\begin{align}
\hat{R}_i = \sum_j \alpha_{ij} R_{i,j}, \quad \hat{R}_i \in \mathbb{R}^c
\end{align}

In contrast to the original region features $R_i$, the attentive region feature mapping $\hat{R}_i$ is more effective to reflect the relevance to the corresponding text representation $H_i$. Then, $\hat{R}_i$ and $H_i$ are further fed into a multi-layer perceptron (MLP) to learn an $f$-dimensional RT embedding as $F_i^{(rt)}$. We denote the whole process of learning RT embedding from image and text pair $(V_i, T_i)$ as follows:

\begin{align}
F_i^{(rt)} = rt(V_i, T_i; \theta_{rt}), \quad F_i^{(rt)} \in \mathbb{R}^f
\end{align}

where function $rt(\cdot)$ denotes the whole networks to obtain RT embedding and $\theta_{rt}$ is the parameters.

### C. Learning Words-Image Embedding

On the other side, the image also has close correlation with some words in the corresponding text. Therefore, the textual attention model is proposed to exploit the correlation between textual words and image to learn the joint representation named WI embedding.

Given an image $V_i$, we first extract the high-level visual features from deep CNNs as $Q_i \in \mathbb{R}^q$, where $q$ is the
dimensionality of visual feature vector. For the corresponding text \( T_i \), each word is embedded and fed into a bidirectional LSTM (Bi-LSTM) to learn semantic word representations as \( E_i \in \mathbb{R}^{e \times l} \), where \( l \) denotes the length of the text and \( e \) denotes the dimensionality of word representation.

To fuse word features \( E_i \) and image features \( Q_i \) for correlation learning, they are first projected into an \( s \)-dimensional common space as follows:

\[
E_i^{(s)} = \tanh(W_e E_i + b_e), \quad E_i^{(s)} \in \mathbb{R}^{s \times l} \tag{7}
\]

\[
Q_i^{(s)} = \tanh(W_q Q_i + b_q), \quad Q_i^{(s)} \in \mathbb{R}^s \tag{8}
\]

where \( W_e \in \mathbb{R}^{s \times e} \) and \( W_q \in \mathbb{R}^{s \times q} \) are the parameter matrices, and \( b_e \) and \( b_q \) are the bias terms. Then, \( E_i^{(s)} \) and \( Q_i^{(s)} \) are fused by the element-wise multiplication:

\[
F_i^{(s)} = E_i^{(s)} \odot (Q_i^{(s)} \cdot 1), \quad F_i^{(s)} \in \mathbb{R}^{s \times l} \tag{9}
\]

where \( 1 \in \mathbb{R}^{s \times l} \) is used to broadcast \( Q_i^{(s)} \) to dimensionality of \( s \times l \). The attention map is then calculated by convolving the fused representation \( F_i^{(s)} \) with kernel length of 1 followed by a softmax function over \( l \) words:

\[
\beta_i = \text{softmax}(\tanh(W_\beta * F_i^{(s)} + b_\beta)), \quad \beta_i \in \mathbb{R}^l \tag{10}
\]

where \( W_\beta \in \mathbb{R}^{s \times l} \) and \( b_\beta \in \mathbb{R}^s \) are the convolutional parameters. Then, the attentive textual features can be calculated as the weighted average of the features over all the words:

\[
\hat{E}_i = \sum_k^l \beta_{i,k} E_{i,k}, \quad \hat{E}_i \in \mathbb{R}^e \tag{11}
\]

In contrast to the original text features \( E_i \), the attentive textual feature mapping \( \hat{E}_i \) is more effective to reflect the relevance to the corresponding image. Then, \( \hat{E}_i \) and \( Q_i \) are further fed into a multi-layer perceptron (MLP) to learn a \( f \)-dimensional WI embedding as \( F_i^{(wi)} \). We denote the whole process of learning WI embedding from image and text pair \((V_i, T_i)\) as follows:

\[
F_i^{(wi)} = wi(V_i, T_i; \theta_{wi}), \quad F_i^{(wi)} \in \mathbb{R}^f \tag{12}
\]

where the function \( wi(\cdot) \) denotes the whole networks to obtain WI embedding and \( \theta_{wi} \) is the parameters.

D. ADVERSARIAL LEARNING OF MULTI-MODAL EMBEDDINGS

The visual and textual attention models have been used to learn RT and WI embeddings respectively. The two types of attention models learn the multi-modal representation based on different types of correlation. The visual attention model focuses on learning the discriminative visual feature, while the textual attention model focuses on learning the image-related word features. Motivated by the recent advance of generative adversarial models [34], [48]–[50], an adversarial learning framework is proposed to reinforce the two attention models mutually to learn more effective multi-modal embeddings.

The generative adversarial networks (GANs) usually contain a generator and a discriminator. In our model, the generator consists of two components for multi-modal embedding generating, i.e., visual attention model and textual attention model. To make the two attention models learn from each other, a classifier \( d(\cdot; \theta_g) \) is built as the discriminator to distinguish the embeddings of RT and WI.

In the learning process, the discriminator is trained to classify the RT and WI embeddings based on the input vectors. The input RT and WI embeddings are assigned with the label of \( \overline{\text{DT}} \). Based on Eq.(6) and Eq.(12), the loss function for the discriminator is formulated as:

\[
L_d(V_i, T_i; \theta_d) = -\log(d(F_i^{(rt)}; \theta_d)) - \log(1 - d(F_i^{(wi)}; \theta_d))
\]

\[
= -\log(d(rt(V_i, T_i; \theta_{rt}); \theta_d)) - \log(1 - d(wi(V_i, T_i; \theta_{wi}); \theta_d)) \tag{13}
\]

where \( \theta_d \) and \( \theta_{wi} \) are kept unchanged during the training of the discriminator.

On the other side, the two attention models (generator) aim to confuse the discriminator. Thus the input RT and WI embeddings are assigned with the label of \( \overline{\text{T0}} \). The generator can be trained to reduce the following loss:

\[
L_g(V_i, T_i; \theta_{gt}) = -\log(1 - d(F_i^{(rt)}; \theta_{gt})) - \log(d(F_i^{(wi)}; \theta_{gt}))
\]

\[
= -\log(1 - d(rt(V_i, T_i; \theta_{rt}); \theta_{gt})) - \log(d(wi(V_i, T_i; \theta_{wi}); \theta_{gt})) \tag{14}
\]

where \( \theta_g \) is kept unchanged when training the generator.

However, as stated by [49], [51], the Jenson-Shannon divergence optimized by the original GANs results in instability problem. To ease the training difficulty of the original GAN, WGAN [49] is proposed with an efficient approximation of the Wasserstein distance (also known as the Earth Mover distance). While the original GANs are driven by a classifier separating real from fake samples, WGANs are driven by a distance measure which quantifies the similarity of two distributions. Therefore, in our experiments, WGAN is adopted to perform the adversarial learning of multi-modal embeddings. Based on Eq.(14) and Eq.(13), the loss of the generator and discriminator can be rewritten respectively as:

\[
L'_g(V_i, T_i; \theta_{rt}, \theta_{gt}) = d(rt(V_i, T_i; \theta_{rt}); \theta_{gt})) - d(wi(V_i, T_i; \theta_{wi}); \theta_{gt}) \tag{15}
\]

\[
L'_d(V_i, T_i; \theta_d) = -d(rt(V_i, T_i; \theta_{rt}); \theta_d) + d(wi(V_i, T_i; \theta_{wi}); \theta_d) \tag{16}
\]

The loss function (15) and (16) are very similar to that of the original GAN, but has better convergence properties.

E. LEARNING MULTI-MODAL EMBEDDINGS FOR MATCHING

Next, the embeddings of RT and WI are combined to learn the image-text matching score. In particular, we build a multi-layer perceptron (MLP) with sigmoid activation in the last
layer to exploit the two types of correlation to learn the matching score. Similar to [52], the pairwise margin ranking loss is used to rank the matching scores of the matched image-text pair \((V_i, T_i)\) and the hardest negative pair \((V_i, T_i^-)\) as follows:

\[
L_m(V_i, T_i, T_i^-; \theta_m, \theta_r, \theta_w) = \max[0, M - m(rt(V_i, T_i; \theta_r), wi(V_i, T_i; \theta_w); \theta_m)] + m(rt(V_i, T_i^-; \theta_r), wi(V_i, T_i^-; \theta_w); \theta_m)]
\]

(17)

where function \(m(\cdot; \theta_m)\) denotes the MLP to calculate the matching score, and \(M\) is the margin parameter. We have also tested the symmetric ranking loss, but the hardest negative scheme performs much better.

To combine the adversarial learning method to learn the multi-modal representation for image-text matching, the loss functions Eq.(15) and Eq.(17) are simultaneously minimized as a overall loss:

\[
L_m(V; \theta_m, \theta_r, \theta_w) = \sum_{i=1}^n L_m(V_i, T_i, T_i^-; \theta_m, \theta_r, \theta_w) + \lambda L'_m(V_i; \theta_m, \theta_r, \theta_w)
\]

(18)

where \(\lambda\) is a hyper-parameter to regulate the importance of the adversarial learning. Note that here we use MLP for matching learning (Eq.17) and attention models for embedding learning (Eq.15). However, most of other matching methods such as DAN [22] and SCAN [43] employ cosine similarity to calculate matching scores. If we use cosine similarity here, the attention models need to undertake these two tasks of matching learning and embedding learning. Since the adversarial networks is also built on the embeddings from the attention models, it is more appropriate to divide these two learning tasks.

Similarly, the adversarial discriminative loss Eq.(16) for training is written as follows:

\[
L_d(V; \theta_d) = \sum_{i=1}^n \lambda L'_d(V_i; \theta_d)
\]

(19)

The training of the proposed model AAMEL can be implemented by matching learning and discriminating learning alternatively with back-propagation. For matching learning, the loss Eq.(18) is minimized to update the parameters \(\theta_m, \theta_r, \text{ and } \theta_w\). For discrimination learning, the parameter \(\theta_d\) is updated according to the loss Eq.(19). In summary, the detailed training procedure is illustrated in Algorithm 1.

After training, the matching score of a given image \(x\) and text \(y\) can be inferred as follows:

\[
\text{score}(x, y) = m(rt(x, y), wi(x, y))
\]

(20)

The time-complexity of the proposed model at inference time is faster than most representative previous works based on attention, like DAN [22] and SCAN [43]. For an image-text pair, our method yields \(O(RD + TD)\), where \(R\) = number of image regions, \(T\) = number of words, and \(D\) = embedding size. SCAN is \(O(R^2TD)\) or \(O(R^2T^2D)\). DAN is \(O(KRD + KTD)\), where \(K\) is the number of pre-defined reasoning steps. As for methods not based on attention mechanism such as DCCA [14], the inference process of these methods is faster. However, the performance of these methods is much worse than that of attention-based methods.

### IV. EXPERIMENTS

In this section, extensive experiments are conducted to analyze the effectiveness of the proposed AAMEL.

#### A. DATASETS

The experiments are conducted on two benchmark datasets, i.e., Flickr30K and MSCOCO, which are collected from Flickr. The details are described below:

- **Flickr30K** [53] is an image-description dataset with 31,783 images containing people and animals. Each image is associated with five sentences written by native annotators. Following [54], 29,783, 1,000, and 1,000 images are used for training, validation, and testing respectively.

- **MSCOCO** [55] contains 82,783 images for training and 40,504 images for validation. Each image is annotated with five sentences. Following [54], the dataset is randomly re-split with 82,783, 4,000, and 1,000 (or 5,000) images for training, validation, and testing respectively.

Then we follow [52] to put the remaining images that originally belong to the validation set into the training set. We present experimental results on both the 1,000 and 5,000 testing images.

#### B. EXPERIMENTAL SETTINGS

For the images, we employ two types of deep CNNs, i.e., VGG19 [61] and ResNet152 [62], pre-trained on ImageNet dataset [63] to extract visual features. For VGG19,
the output of “conv5-4” is used as the region features with the dimension of $14 \times 14 \times 512$ for the visual attention model. And the output of the layer “fc7” is used as the image features with the dimension of 4096 for the textual attention model. For ResNet152, the spatial features before the average pooling layer is used as the region features with the dimension of $7 \times 7 \times 512$ for the visual attention model. The output of the average pooling layer is used as the image features with the dimension of 2048 for the textual attention model. The texts, we employ 300-dimensional pre-trained GloVe [64] representations to embed each word. The LSTM and Bi-LSTM are set to have 512 and 256 neurons respectively. The MLPs in the visual and textual attention models are all set with the network structure 2048-1024-1024 activated with \textit{tanh} function. The MLP built for matching score learning is set with structure 2048(\textit{tanh})-1024(\textit{tanh})-1(\text{sigmoid}). The margin in Eq.(17) is experimentally set to be 0.1. The balance parameter \(\lambda\) in Eq.(18) is set to be 0.2 for relatively good performance. m-steps and d-step in Algorithm 1 are set to be 1 and 4 respectively for relatively fast and stable convergence.

The implementation is conducted using Pytorch\footnote{https://github.com/pytorch/pytorch} deep learning tool on four GPUs of NVIDIA GTX 1080Ti. We use Adam optimizer [65] with an initial learning rate of 0.0002 to train for 15 epochs. Then we fine-tune the model with a lowered learning rate of 0.00002 for another 15 epochs.

### TABLE 1. Cross-modal retrieval results on dataset Flickr30k.

| Methods          | Sentence Retrieval | Image Retrieval |
|------------------|--------------------|-----------------|
|                  | R@1 | R@5 | R@10 | Med r      | R@1 | R@5 | R@10 | Med r      |
| DCCA (VGG) [14]  | 27.9 | 56.9 | 68.2 | 4.0        | 26.8 | 52.9 | 66.9 | 4.0        |
| CCA (VGG+FV) [35]| 33.3 | 62.0 | 74.7 | 3.0        | 25.6 | 53.2 | 66.8 | 5.0        |
| sm-LSTM (VGG) [15]| 42.5 | 71.9 | 81.5 | 2.0        | 30.2 | 60.4 | 72.3 | 3.0        |
| 2WayNet (VGG) [56]| 49.8 | 67.5 | -   | -          | 36.0 | 55.6 | -   | -          |
| BSSAN (VGG) [57] | 44.6 | 74.9 | 84.3 | 2.0        | 33.2 | 62.6 | 72.9 | 3.0        |
| VSE++ (ResNet) [52]| 52.9 | 79.1 | 87.2 | 1.0        | 39.4 | 69.6 | 79.1 | 2.0        |
| DPC (ResNet) [58] | 55.6 | 81.9 | 89.5 | 1.0        | 39.1 | 69.2 | 80.9 | 2.0        |
| DAN (ResNet) [22] | 55.0 | 81.8 | 89.0 | -          | 39.4 | 69.2 | 79.1 | -          |
| SCO (ResNet) [59] | 55.5 | 82.0 | 89.3 | -          | 41.1 | 70.5 | 80.1 | -          |
| SCAN (ResNet)  [43]| 67.4 | 90.3 | 95.8 | -          | 48.6 | 77.7 | 85.2 | -          |
| AAMEL (VGG)       | 54.2 | 79.0 | 87.6 | 1.0        | 38.4 | 68.6 | 78.0 | 2.0        |
| AAMEL (ResNet)    | 68.5 | 91.2 | 95.9 | 1.0        | 49.7 | 79.2 | 86.4 | 2.0        |

### TABLE 2. Cross-modal retrieval results on dataset MSCOCO.

| Methods          | Sentence Retrieval | Image Retrieval |
|------------------|--------------------|-----------------|
|                  | R@1 | R@5 | R@10 | Med r      | R@1 | R@5 | R@10 | Med r      |
| CCA (VGG+FV) [35]| 38.9 | 68.4 | 80.1 | 2.0        | 25.6 | 60.4 | 76.8 | 4.0        |
| Order-embeddings (VGG) [60]| 46.7 | 88.9 | 2.0 | 38.9 | -   | 85.9 | 2.0        |
| sm-LSTM (VGG) [15] | 53.2 | 83.1 | 91.5 | 1.0        | 40.7 | 75.8 | 87.4 | 2.0        |
| 2WayNet (VGG) [56]  | 55.8 | 75.2 | -   | -          | 39.7 | 63.3 | -   | -          |
| BSSAN (VGG) [57]   | 56.0 | 82.6 | 91.3 | 1.0        | 41.7 | 76.7 | 88.5 | 2.0        |
| VSE++ (ResNet) [52]| 64.6 | 89.1 | 95.7 | 1.0        | 52.0 | 83.1 | 92.0 | 1.0        |
| DPC (ResNet) [58]  | 65.6 | 89.8 | 95.5 | 1.0        | 47.1 | 79.9 | 90.0 | 2.0        |
| CHAIN-VSE (ResNet) [16]| 61.2 | 89.3 | 95.8 | 1.0        | 46.6 | 81.9 | 90.9 | 2.0        |
| SCO (ResNet) [59]  | 69.9 | 92.9 | 97.3 | -          | 56.7 | 87.5 | 94.8 | -          |
| GXN (ResNet) [44]  | 68.5 | 97.9 | -   | 56.6 | -   | 94.5 | 1.0        |
| SCAN (ResNet) [43] | 72.7 | 94.8 | 98.4 | -          | 58.8 | 88.4 | 94.8 | -          |
| AAMEL (VGG)       | 63.1 | 87.5 | 94.1 | 1.0        | 49.9 | 82.5 | 90.8 | 2.0        |
| AAMEL (ResNet)    | 74.3 | 95.4 | 98.2 | 1.0        | 59.9 | 89.0 | 95.1 | 1.0        |

### Table -- Cross-modal retrieval results on dataset MSCOCO.

| Methods          | Sentence Retrieval | Image Retrieval |
|------------------|--------------------|-----------------|
|                  | R@1 | R@5 | R@10 | Med r      | R@1 | R@5 | R@10 | Med r      |
| Order-embeddings (VGG) [60]| 23.3 | 65.0 | 5.0 | 18.0 | -   | 57.6 | 7.0        |
| VSE++ (ResNet) [52]| 41.3 | 69.2 | 81.2 | 2.0        | 30.3 | 59.1 | 72.4 | 4.0        |
| DPC (ResNet) [58]  | 41.2 | 70.5 | 81.1 | 2.0        | 25.3 | 53.4 | 66.4 | 5.0        |
| SCO (ResNet) [59]  | 42.8 | 72.3 | 83.0 | -          | 33.1 | 62.9 | 75.5 | -          |
| GXN (ResNet) [44]  | 42.0 | 84.7 | 2.0 | 31.7 | -   | 74.6 | 3.0        |
| SCAN (ResNet) [43] | 30.4 | 82.2 | 90.0 | -          | 38.6 | 69.3 | 80.4 | -          |
| AAMEL (VGG)       | 40.2 | 67.7 | 78.5 | 2.0        | 24.6 | 55.9 | 65.1 | 5.0        |
| AAMEL (ResNet)    | 51.9 | 84.2 | 91.2 | 1.0        | 39.9 | 71.3 | 81.7 | 3.0        |
Dropout technique [66] is employed for MLPs with a probability of 0.5 to reduce over-fitting during the training process. Batch Normalization [67] is also employed to reduce the occurrence of covariate shift within the neural networks. To evaluate our method on the testing set, we choose the snapshot of the trained model which performs best on the validation set.

C. COMPARED METHODS

Our approach is compared with the state-of-the-art methods on the tasks of image-text and text-image retrieval. The baselines include: CCA-based methods DCCA [14] and CCA (FV GMM + HGLMM) [35], ranking based methods sm-LSTM (VGG) [15], Order-embeddings (VGG) [60], VSE++ (ResNet) [52], BSSAN (VGG) [57], DPC (ResNet) [58], DAN (ResNet) [22], CHAIN-VSE (ResNet) [16], SCO (ResNet) [59], GXN (ResNet) [44], and SCAN (ResNet) [43], and other methods 2WayNet (VGG) [56].

D. EXPERIMENTAL RESULTS

In this section, the experimental results are presented and analyzed to evaluate the proposed approach for cross-modal retrieval on two datasets. R@K (with K = 1, 5, 10) is adopted as the evaluation metric. It represents the percentage of queries in which at least one correct match is ranked among the top K matches. The metric of median rank (Med r) of the top-ranked correct result is also employed. The performance of AAMEL and the compared methods on Flickr30K and MSCOCO are presented in Table 1 and Table 2 respectively.

From Table 1, it can be seen that both our two methods of AAMEL (VGG) and AAMEL (ResNet) consistently outperform the compared methods with corresponding pre-trained CNNs. CCA-based methods such as DCCA [14] and CCA [35] show relatively low performance because the generalized eigenvalue problem cannot be well solved by stochastic gradient descent [37]. DAN [22], SCO [59], and SCAN [43] obtain the state-of-the-art results for cross-modal search. Compared to DAN and SCAN, our method AAMEL (ResNet) improves the performance consistently on all metrics. Since both DAN and SCAN have employed the attention mechanism for matching learning, the improvement validates the effectiveness of adversarial learning for image-text matching. Figure 3 illustrates several qualitative results of descriptions retrieved by the given images on Flickr30K. We present top-5 ranked descriptions based on the matching scores calculated by our approach for each sample. Groundtruth sentences are checked and false ones are crossed. One can see that the retrieved sentences can well describe the given images.

The results on dataset MSCOCO are presented in Table 2. The several rows above are the results tested on the 1K test images and their associated sentences. It can be seen that our model AAMEL achieves improvement on various metrics over the compared baselines in most cases. SCAN [43] is the state-of-the-art method which integrates image and text adversarial generative models into image-text searching. Our method outperforms SCAN except on R@10 in image-to-text retrieval, which confirms that the proposed model is effective to learn the multi-modal embeddings for image-text matching. We also make tests on all the 5K images and the corresponding descriptions, and present the comparison results in the rows below. It can be observed that the proposed method still achieves the best performance, which again demonstrates its effectiveness. Figure 4 shows the qualitative results of retrieved images given description queries on MSCOCO. Each description corresponds to a groundtruth image.

![FIGURE 3. Qualitative results of descriptions retrieved by our model given four image queries on Flickr30K. Groundtruth matched sentences are marked with ticks, while false ones are marked with crosses. Although some sentences such as (b)2 and (c)4 are regarded as incorrect results, these descriptions share similar meanings which are reasonable mismatches.](image-url)
A group of young children kicking a soccer ball around a field.

Two people riding down the side of a ski slope.

A banana that is sitting on a plate.

FIGURE 4. Qualitative results of images retrieved by our model given three descriptions on MSCOCO. For each description, we present top-5 ranked images from left to right. We outline the correct matches in green boxes and false matches in red boxes. Note that most incorrect matches predicted by our model are also reasonable results.

FIGURE 5. Four examples of image-text pairs for attention visualization. For each pair, we present both the visualization of attended regions in the image and attended words in the corresponding descriptions. The regions and words focused by our model are drawn with red colors.

For each description, we present top-5 ranked images from left to right. The groundtruth and false matches are outlined in green and red boxes respectively.

E. ABLATION STUDY

To analyze the contributions of different components of AAMEL, we ablate the proposed model and demonstrate the effectiveness of each component: 1) Visual attention model to learn RT embedding (VA); 2) Textual attention model to learn WI embedding (TA); 3) Adversarial Networks to learn the multi-modal embeddings (AN).

Table 3 shows the results of the ablative experiments on the dataset Flickr30k. V and T are the simple versions of VA and TA respectively by changing the two kinds of
VOLUME 8, 2020

description. On the other side, the important semantic words These regions are closely related to the corresponding text attention on the regions of the little girl with the basket. The first image-text pair shows that AAMEL draws the visual image regions and textual words correctly. For example, the results demonstrate that the proposed method outperforms state-of-the-art baselines. Ablation study also verifies the effectiveness of the main components of AAMEL.

F. VISUALIZATION OF ATTENTIONS

To further analyze the effectiveness of the visual and textual attention models in AAMEL, we visualize the attentions learned by the model. Upsample and Gaussian filter are used to visualize the attention weights of images and red stroking is used to highlight the attended words. Figure 5 presents several examples of the attended image and text pairs in Flickr30K. From the results, one can see that the two attention models can complement each other for matching. With the adversarial networks, VA+TA+AN (the full model) further improves the performance with about 4% on different metrics. It validates that adversarial learning reinforces the two attention models to learn more effective embeddings.

G. PARAMETER SENSITIVITY

To evaluate how the balance parameter $\lambda$ affects the performance of AAMEL, the sensitivity experiments are conducted for different values of $\lambda$. Figure 6 shows the performance (R@1) of AAMEL on Flickr30k. When $\lambda = 0$, the training is governed by the attention models, which leads to that the attention models are not effective to learn from each other without the adversarial networks. When $\lambda$ becomes larger, the model concentrates more on the adversarial learning. However, too large values of $\lambda$ also influence the learning of the attention models. It can be seen that AAMEL obtains the best performance at about $\lambda = 0.2$.

V. CONCLUSION

In this paper, we propose a novel deep method named Adversarial Attentive Multi-modal Embedding Learning (AAMEL) for image-text matching, which engages two processes in an interplay game. To capture the correlation effectively, two attention models are proposed to learn the two types of correlation between image and text. Then the adversarial learning method is proposed to reinforce the attention models mutually to make the learned multi-modal representation more effectively for image-text matching. We employ two benchmark datasets to conduct experiments on the tasks of image-to-text and text-to-image retrieval. The experimental results demonstrate that the proposed method outperforms state-of-the-art baselines. Ablation study also verifies the effectiveness of the main components of AAMEL.

The main limitation of the proposed method is that the inference process of AAMEL is slower than traditional methods based on CCA or hash. In the future, we will explore how to integrate different levels of text and image features to learn more effective multi-modal embeddings for image-text matching. Furthermore, we plan to generalize our method to other multimodal data such as video and music for cross-modal retrieval.

REFERENCES

[1] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, “StructCap: Structured semantic embedding for image captioning,” in Proc. ACM Multimedia Conf. (MM), Mountain View, CA, USA, Q. Liu, R. Lienhart, H. Wang, S. K. Chen, S. Boll, Y. P. Chen, G. Friedland, J. Li, and S. Yan, Eds., Oct. 2017, pp. 46–54.

[2] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, “Image captioning with both object and scene information,” in Proc. ACM Conf. Multimedia Conf. (MM), Amsterdam, The Netherlands, A. Hanjalic, C. Snoek, M. Worringer, D. C. A. Balterman, B. Huet, A. Kelliker, Y. Kompatakis, and J. Li, Eds., Oct. 2016, pp. 1107–1110.

[3] C. Wang, H. Yang, C. Bartz, and C. Meinel, “Image captioning with deep bidirectional LSTMs,” in Proc. ACM Conf. Multimedia Conf. (MM), Amsterdam, The Netherlands, A. Hanjalic, C. Snoek, M. Worringer, D. C. A. Balterman, B. Huet, A. Kelliker, Y. Kompatakis, and J. Li, Eds., Oct. 2016, pp. 988–997.
[4] K. Fu, J. Li, J. Jin, and C. Zhang, “Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 12, pp. 5910–5921, Dec. 2018.

[5] J. Dong, X. Li, and C. G. M. Snoek, “Predicting visual features from text for image and video caption retrieval,” IEEE Trans. Multimedia, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.

[6] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, “GLA: Global–local attention for image description,” IEEE Trans. Multimedia, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[7] J. Li, J. Z. Wang, and G. Wiederhold, “IRM: Integrated region matching for image retrieval,” in Proc. 8th ACM Int. Conf. Multimedia, Los Angeles, CA, USA, S. Ghandeharizadeh, S. Chang, S. Fischer, J. A. Konstan, and K. Nahrstedt, Eds., Oct./Nov. 2000, pp. 147–156.

[8] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, “Optimal coding for large scale partial-duplicate Web image search,” in Proc. 18th Int. Conf. Multimedia, Firenze, Italy, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds., Oct. 2010, pp. 511–520.

[9] M. Soleymani, M. Riegler, and P. Halvorsen, “Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content,” in Proc. Int. Conf. Conf. Multimedia Retr. (ICMR), Budapest, Roman, J. Jonescu, N. Sisca, J. Meng, M. Larson, R. Lienhart, and C. Snoek, Eds., Jun. 2017, pp. 251–259.

[10] F. Shen, Y. Yang, L. Liu, W. Liu, D. Tao, and H. T. Shen, “Asymmetric binary coding for image search,” IEEE Trans. Multimedia, vol. 19, no. 9, pp. 2022–2032, Sep. 2017.

[11] Y.-G. Jiang, J. Wang, X. Xue, and S.-F. Chang, “Query-adaptive image and visual content,” in Proc. 18th Int. Conf. Multimedia, Firenze, Italy, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds., Oct. 2010, pp. 511–520.

[12] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, “Discriminative binary coding for image search,” in Proc. 17th Int ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), Gold Coast, QLD, Australia, S. Geva, A. Trotman, P. Brusz, C. L. A. Clarke, and K. Järvelin, Eds. Jun. 2014, pp. 395–404.

[13] L. Ma, Z. Lu, L. Shang, and H. Li, “Multimodal convolutional neural networks for matching image and sentence,” in Proc. Proc. Int. Conf. Multimedia Vis. (ICMV), Santiago, Chile, Yang, H. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3441–3450.

[14] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 7254–7262.

[15] J. Wehrmahn and R. C. Barros, “Bidirectional retrieval made simple,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 7718–7726.

[16] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in Proc. Proc. Comput. Vis. 15th Eur. Conf. (ECCV) ( Lecture Notes in Computer Science), Munich, Germany, vol. 11205, V. Ferrari, M. Hebert, C. Falquez, J. Chang, and Y. Li, Eds. Sep. 2018, pp. 707–723, doi: 10.1007/978-3-030-12364-5_42.

[17] Y. Huang, Y. Li, Y. Yin, and Y. Zhao, “Deep metric learning for cross-modal matching,” IEEE Trans. Multimedia, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” Neural Comput., vol. 16, no. 11, pp. 2664–2684, Dec. 2004.

[19] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in Proc. 9th Int. Conf. Mach. Learn. (ICML), Atlanta, GA, USA, Jun. 20, 2013, pp. 1247–1255. [Online]. Available: http://proceedings.mlr.press/v28/andrew13.html

[20] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in Proc. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 5005–5013.

[21] H. Jiang, J.-W. Hu, and D. C. Liu, “Dual attention networks for multimodal reasoning and matching,” in Proc. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2156–2164.

[22] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, “FVQA: Fact-based visual question answering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 10, pp. 2413–2427, Oct. 2018.

[23] Y. Liu, X. Zhang, F. Huang, X. Tang, and Z. Li, “Visual question answering via attention-based syntactic structure tree-LSTM,” Appl. Soft Comput., vol. 82, Sep. 2019, Art. no. 105584.
