Robust Multimodal Representation Learning with Evolutionary Adversarial Attention Networks

Feiran Huang, Alireza Jolfaei, Ali Kashif Bashir

***Abstract*—Multimodal representation learning is beneficial for many multimedia-oriented applications such as social image recognition and visual question answering. The different modalities of the same instance (e.g., a social image and its corresponding description) are usually correlational and complementary. Most existing approaches for multimodal representation learning are not effective to model the deep correlation between different modalities. Moreover, it is difficult for these approaches to deal with the noise within social images. In this paper, we propose a deep learning-based approach named Evolutionary Adversarial Attention Networks (EAAN), which combines the attention mechanism with adversarial networks through evolutionary training, for robust multimodal representation learning. Specifically, a two-branch visual-textual attention model is proposed to correlate visual and textual content for joint representation. Then adversarial networks are employed to impose regularization upon the representation by matching its posterior distribution to the given priors. Finally, the attention model and adversarial networks are integrated into an evolutionary training framework for robust multimodal representation learning. Extensive experiments have been conducted on four real-world datasets, including PASCAL, MIR, CLEF, and NUS-WIDE. Substantial performance improvements on the tasks of image classification and tag recommendation demonstrate the superiority of the proposed approach.**

***Index Terms*—Adversarial networks, attention model, evolutionary, multimodal, representation learning.**

I. Introduction

With the advent of the Internet, multimodal data has become increasingly popular in the everyday life of people in the past few years. People share photos, write comments, and watch videos on various Internet sites such as Facebook, Twitter, and Flickr. Different modalities of multimodal data usually carry correlational and complementary information. Learning a multimodal representation to transform multiple modalities into a joint vector is very useful to extract the feature needed for further analysis and applications. The learned representation has been extensively applied to multimedia-related tasks such as social image classification [1], [2], [3], tag recommendation [4], [5], and visual question answer [6], [7]. In the field of social multi-media, the representation learning of multimodal data is becoming increasingly important and has also attracted growing research interests.

However, the representation learning of multimodal data also brings some tough challenges to researchers. First, there are various manifestations of social images such as visual content and textual captions. These modalities are characterized by different statistical properties and exist in heterogeneous feature spaces. Therefore, the representation learning approaches should fuse different modalities by effectively bridging the modal gap. Second, different modalities usually carry complementary information from each other. It is necessary to extract comprehensive and non-redundant features from the input multimodal data. Third, since social images are shared freely and the corresponding descriptions are written casually, a lot of noisy information may exist in these multimodal content. Therefore, the approaches should learn robust representation to deal with the noise within multimodal data.

In the past few years, there have been a lot of approaches on multimodal representation learning. These methods can be generally classified into two types. The first line of research aims to transform multiple modalities of input data into a joint embedding vector. Ngiam *et al.* [8] proposed a bimodal deep denoising autoencoder to encode unlabeled data for multimodal representation. Srivastava *et al.* [9] employed a deep boltzmann machine to fuse multiple data modalities for representation learning. However, the correlation between different modalities is not fully mined by these methods. The second strategy of multimodal representation learning projects different modalities to a shared vector space with a constraint to capture the cross-modal correlation. DCCA [10], [11] is an extension of Canonical Correlation Analysis with deep learning techniques which learns projection of two views by maximizing cross-view relations. WSABIE [12] and DeVISE [13] both employed a hinge loss to rank the similarity of input images-text pairs and project them into separate vectors with the same embedding size. These approaches are usually capable of excavating the cross-modal interaction, but they are not effective to capture the complementary information from different modalities. Moreover, It is difficult for the two types of approaches to deal with the noise within social images for robust representation learning without additional constraints.

On the other side, clues can also be found from the correlated and sequential characteristics of social images to learn multimodal representation. First, there exists fine-grained correlation between different modalities of images and texts. Take Figure 1 as an example. It can be seen that some specific

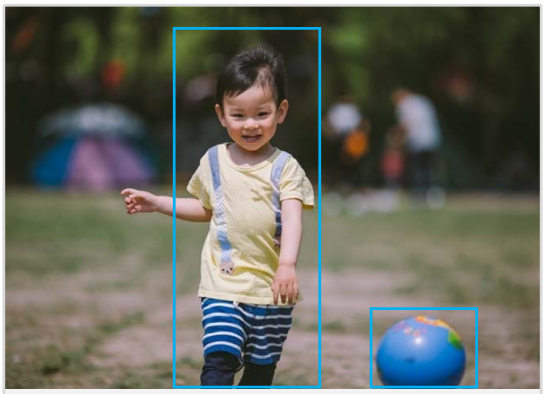
This work was supported in part by the National Natural Science Foundation of China (Grant No. 61906075, 62002068, 61932010, 61932011, 61972178), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011920, 2019A1515011276, 2019A1515011753), Guangdong Provincial Key R&D Plan (Grant No. 2019B1515120010, 202020022911500032, 2019B010136003), Science and Technology Program of Guangzhou, China (Grant No. 202007040004).

Feiran Huang is with College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou, 510632, China. (e-mail: huangfr@ieee.org)

Alireza Jolfaei is with Department of Computing, Macquarie University, Sydney, Australia.

Ali Kashif Bashir is with Department of Computing and Mathematics, Manchester Metropolitan University, UK.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEVC.2021.3066285, IEEE Transactions on Evolutionary Computation



Description

The smiling boy in a yellow T-shirt is playing with a ball at park Nikon D850.

Fig. 1: An example of social image and its description: (1) the image and its corresponding description fine-granularly correlated to each other. The words in brown boxes can be easily found in the image, and the image areas of "ball" and "boy" cover the main semantics of the corresponding texts. (2) There exists a lot of noise in the social multimodal data. For example, "Nikon D850" is the camera model taking this picture, yet it has no contribute to the representation learning.

words (e.g., "smiling" and "boy") are more relevant to the image. For another, the visual areas of "ball" and "boy" cover more semantic information from the corresponding sentence than other regions. If the fine-grained correlation can be well parsed, the images and texts are easier to be modeled to mine the complementary information within multi-modalities. Attention mechanism has been extensively used as an effective technique to learn salient features. It has been successfully applied to many vision and language-related tasks, such as visual question answering [6], [7], image captioning [14], [15], and cross-modal retrieval [16], [17]. However, employing the attention mechanism for multimodal representation learning still needs further study. Second, since the multi-modalities of social images are filled with noisy information as shown in Figure 1, it is necessary to model the uncertainty within social images during the representation learning process. Generative adversarial networks (GANs) have emerged recently as a powerful generative learning approach to model the distribution of data, which benefit many tasks, such as image generation and style transfer. It has been demonstrated that well-designed adversarial networks are effective to learn representation for image [18], [19] and text [20]. However, these models are not devised for multimodal representation learning against noise.

To deal with the challenges, we propose a novel approach named Evolutionary Adversarial Attention Networks (EAAN) for robust multimodal representation learning. Specifically, a two-branch visual-textual attention model is proposed to correlate the modalities of image and text fine-granularly. To make the learned representation more effective, siamese similarity with an asymmetrical attention strategy is employed to guide the learning of attention weights. Then adversarial networks are employed to constrain the representation by matching its posterior distribution to the given priors. The adversarial learning model acts as a regularizer that regulates the representation more robust to deal with noise. The adversarial learning reinforces the learned joint multimodal representation more robust to deal with noisy information. Finally, the attention model and adversarial networks are integrated into an evolutionary training framework for robust multimodal representation learning. The contributions of this work can be summarized as follows:

- We investigate the problem of learning multimodal representation by excavating the fine-grained correlation and modeling the uncertainty to deal with noise. Our model is unsupervised and task independent, which is suitable for multiple types of data mining tasks.
- We propose a novel approach named Evolutionary Adversarial Attention Networks (EAAN) for joint multimodal representation learning. To the best of our knowledge, we are the first to learn multimodal representation combining attention mechanism and adversarial learning with evolutionary training.
- Extensive experiments have been carried out on four real-world datasets. The proposed approach makes significant improvement of performance over state-of-the-art methods for multimodal representation learning.

This paper improves its preliminary version [21] in terms of both experimental performance and technique. First, we design a two-branch visual-textual attention model to learn more effective representation and employ an asymmetrical attention strategy to learn the attention weights. Second, we employ WGAN [22] to replace the original GAN to make the training more stable. Third, evolutionary algorithm is employed during the training process to select the hyper-parameters more effectively. Fourth, more extensive experiments are conducted and one more dataset is added to evaluate our method. Finally, the proposed approach and model settings are presented and described in more detail.

The remainder of the paper proceeds as follows. Related work is reviewed in Section II. Next, we define the studied problem and introduce the proposed EAAN in detail. The experiments are then elaborated in Section IV. Lastly, we draw conclusions.

II. RELATED WORK

A. Multimodal Representation Learning

The representation learning of unimodal data has been broadly studied [23], [24]. With the explosive growth of social media, increasing interests have been drawn on the multimodal representation learning. It supports many applications, such as social image classification [25], visual question answering [6], [7], and image captioning[14], [26], [15].

At early stage, many statistics-based multimodal representation learning methods [27], [28], [29] have been proposed. Blei *et al.* [27] proposed Corr-LDA to model the joint distribution of multimodal data with corresponding latent Dirichlet allocation, in which multiple conditional relations between the

representations of images and texts are found. Rasiwasia *et al.* [29] combined semantic abstraction with the encoding of cross-modal relations to learn representations for retrieval task. Though these methods have achieved certain performance on multimodal representation learning, it is difficult for them to detect the high-level features with shallow structures.

Recently, many deep neural network-based methods have been proposed for multimodal representation learning. These approaches can be generally classified into two types. The first line of research aims to transform multiple data modalities into a joint embedding vector. The easiest way is to represent each modality separately and then concatenate them as the joint representation. However, simple concatenation of separate representations is easy to result in large length vector which contains redundant information. Ngiam *et al.* [8] proposed a bi-modal deep denoising autoencoder pre-trained with sparse RBMs to learn multimodal features from unlabeled data. Srivastava and Salakhutdinov [9] built a deep boltzmann machine network to fuse multiple data modalities for representation learning. Suk *et al.* [30] employed multimodal DBM to learn feature embedding from 3D patches and then performed Alzheimers disease classification on imaging data. However, These methods are not good at capturing the relationships between different modalities.

The second strategy of multimodal representation learning projects different modalities to a shared vector space with a constrain to capture the cross-modal correlation. Feng *et al.* [31] proposed a deep model named correspondence autoencoder to minimize the correlation loss and representation learning loss jointly for cross-modal retrieval. DCCA [10], [11] is a deep learning-based Canonical Correlation Analysis which learns complex nonlinear projection of two views by maximizing the cross-view relations. WSABIE [12] and DeVISE [13] both employed a hinge ranking loss to learn the transformations of images and texts into a shared representation space. DSPN [32] uses a two-branch network to combine ranking constraints with structure preservation constraints for image-text representation learning. CMDN [33] is proposed to integrate cross-media and intra-media correlation to learn shared representations with a two-stage deep framework. ACMR [34] is proposed to learn modality-independent and discriminative representations with an adversarial structure for cross-modal retrieval. However, ACMR is mainly applied to retrieval task and the adversarial structure is employed to discriminate the image and text. These methods well exploit the cross-modal interactions, but they cannot effectively capture the complementary information from different modalities. Moreover, these two types of methods both suffer from lack of additional constraints to deal with the noise within social images for robust representation learning.

B. Attention Mechanism

Attention mechanism has been extensively introduced to the field of computer vision [6], [14], [26] and natural language processing [35], [36], [37].

In machine translation, Bahdanau *et al*. [35] proposed a neural machine translation model containing a bidirectional RNN as the encoder and an attention layer as the decoder to predict targeting sentence. It can automatically make the alignment of related words in the source sentence with each predicted word. Gehring *et al*. [36] extended the sequence to sequence structure from RNN to CNN for translation, which made a large improvement on training speed. In image classification, Wang *et al*. [38] built a residual attention network with both top-down and bottom-up attention structure to classify images. In image captioning, Xu *et al*. [14] employed two types of attentions, i.e., hard attention and soft attention, to focus on salient objects while generating output sequence. You *et al*. [26] employed a multi-level attention model to semantically select important concepts and visual regions to predict captions. In visual question answering, SANs [6] was proposed to build a stacked attention strategy to predict the answer step by step. Lu *et al*. [7] built a co-attention model to predict answers by jointly reasoning upon image and question attention. Different from existing attention models, for each social image and its textual description, our attention model learns the multimodal embedding by merge these two modalities with deep fusion. This model well explores the fine-grained relation between the image and text by two branch attentions, i.e., visual attention and textual attention. Especially, the visual attention branch is built to capture the alignment between image regions and the description while the textual attention branch is built for the alignment between textual words and the image.

C. Generative Adversarial Networks

Generative adversarial networks (GANs) [39] have been introduced recently as a new technique of modeling distributions of data. The core concept of the GANs is the adversarial training of its two main components, i.e., generator and discriminator. The generator is used to generate fake data from a prior sample while the discriminator tries to differentiate fake data from the real data. With a minimax game, the two networks of generator and discriminator are trained iteratively against one another. Since the original GANs suffer the issues of mode collapse, instability, and low quality, some variants are employed to address these problems. DCGANs [18] and WGANs [22], [40] were proposed to ease the training difficulty and avoid the potential issue of mode collapse. GANs have also been extended to use supervised knowledge. For instance, conditional GANs [41] were constructed to generate images with label information.

Due to the powerful ability of GANs, a number of adversarial training algorithms have been proposed recently for representation learning [18], [19], [20], [42], [43], [44]. Adversarial autoencoder [43] is proposed to learn representation from unlabeled data by matching the posterior of the hidden state of the autoencoder with a prior distribution. Donahue *et al*. [42] proposed bidirectional generative adversarial networks for unsupervised image feature learning by projecting data back into the latent space with an additional encoder. Different from these works, we employ the adversarial learning to make the learned multimodal representation match a prior distribution, which acts as a regularizer for robust representation learning.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEVC.2021.3066285, IEEE Transactions on Evolutionary Computation

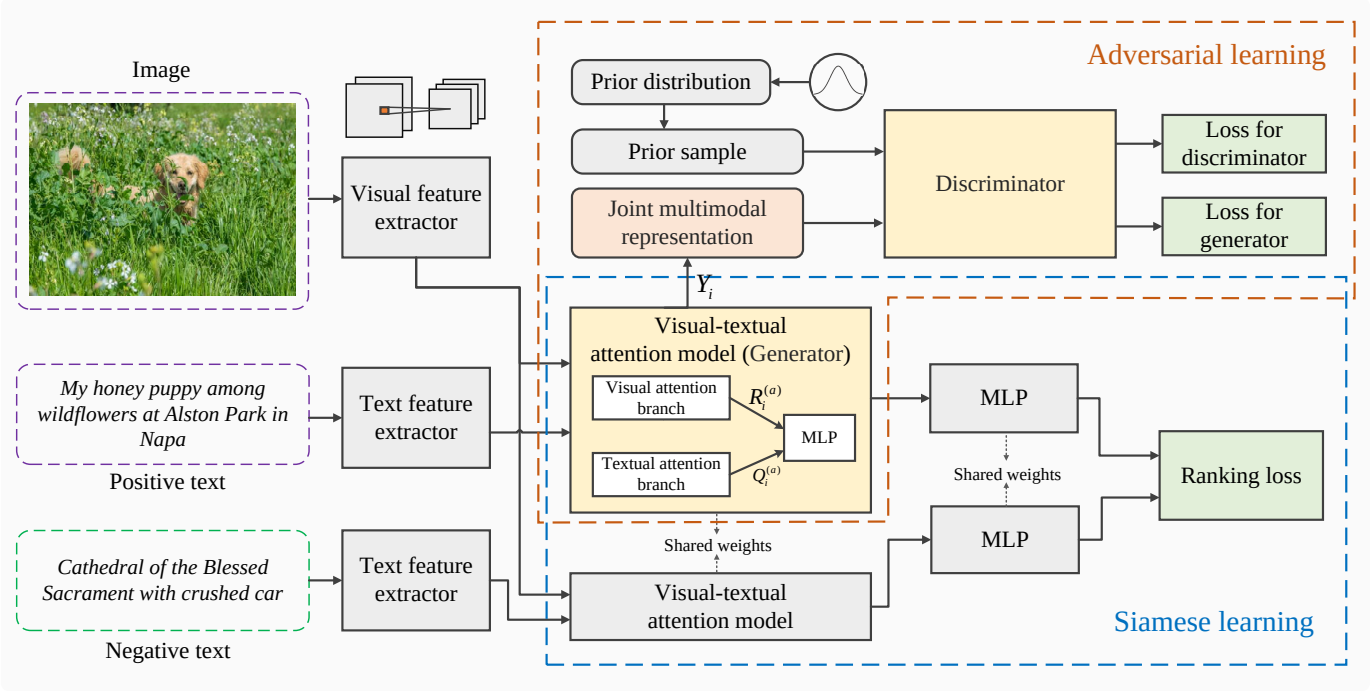


Fig. 2: The framework of EAAN. It mainly contains a siamese learning module and an adversarial learning module.

III. Multimodal Representation Learning

In this section, we first define the studied problem and outline the framework of EAAN. Then we detail the two components of EAAN, i.e., visual-textual attention model and adversarial learning model. Finally, the two components are integrated for joint multimodal representation learning.

A. Problem Statement

Without loss of generality, we consider two of most common modalities, i.e., the image and the textual description. Let $\mathcal{V} = \{V_1, ..., V_i, ..., V_n\}$ and $\mathcal{T} = \{T_1, ..., T_i, ..., T_n\}$ denote a set of n images and the corresponding text descriptions respectively. Then, for each image V_i and text T_i, our goal is to learn a d-dimensional joint representation Z_i. After training, the generated representation can be applied to various tasks e.g., multi-label classification and tag recommendation.

Figure 2 illustrates the framework of EAAN. Specifically, the visual-textual attention model with two branch networks is proposed to learn the multimodal representation by capturing fine-grained cross-modal correlation. The visual attention branch is built to capture the alignment between image regions and the text while the textual attention branch is built for the alignment between textual words and the image. Then siamese similarity with an asymmetrical attending policy is employed to learn the attention weights with pair-wised hinge rank loss. To deal with the noise within social images, adversarial learning is then employed to impose a prior distribution on the learned representation as a regularizer. It makes the generated multimodal representation considering uncertainty via the adversarial training between the discriminator and generator. Through the adversarial process, it is expected that the learned representation is more consistent with the underlying semantics of the raw data with much noisy information. Finally, the attention model and adversarial learning model are integrated into an evolutionary learning framework for robust multimdoal representation learning.

B. Visual-Textual Attention Model

As aforementioned, there exist two types of fine-granularly cross-modal correlations as shown in Figure 1. Some visual regions cover more semantics from the corresponding text while some specific words are more relevant to the image. We propose a two-branch visual-textual attention model to excavate these two types of correlations for multimodal representation learning. The neural structure of our two-branch attention model is shown in Figure 3.

1) Visual attention branch: Given an image-text pair (V_i, T_i), our aim is to discover the salient visual region features most related to the corresponding description. For image V_i, we use pretrained deep convolutional neural networks to extract the region features $R_i \in \mathbb{R}^{e \times m \times m}$, where e is the dimension of each region and $m \times m$ is the number of regions. As for the text T_i, we embed each word as pretrained word embeddings and feed the sequence to an LSTM. Then we use the last cell's output as text features $H_i \in \mathbb{R}^h$.

To correlate the region features R_i and text features H_i, we first transform these them into a common vector space and then fuse them with element-wise multiplication as

$$R_i' = tanh(W_r R_i + b_r), \quad R_i' \in \mathbb{R}^{c \times m \times m}, \tag{1}$$

$$H_i' = tanh(W_h H_i + b_h) \cdot \mathbf{1}, \quad H_i' \in \mathbb{R}^{c \times m \times m}, \tag{2}$$

$$F_i^{(c)} = R_i' \odot H_i', \quad F_i \in \mathbb{R}^{c \times m \times m}, \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEVC.2021.3066285, IEEE Transactions on Evolutionary Computation

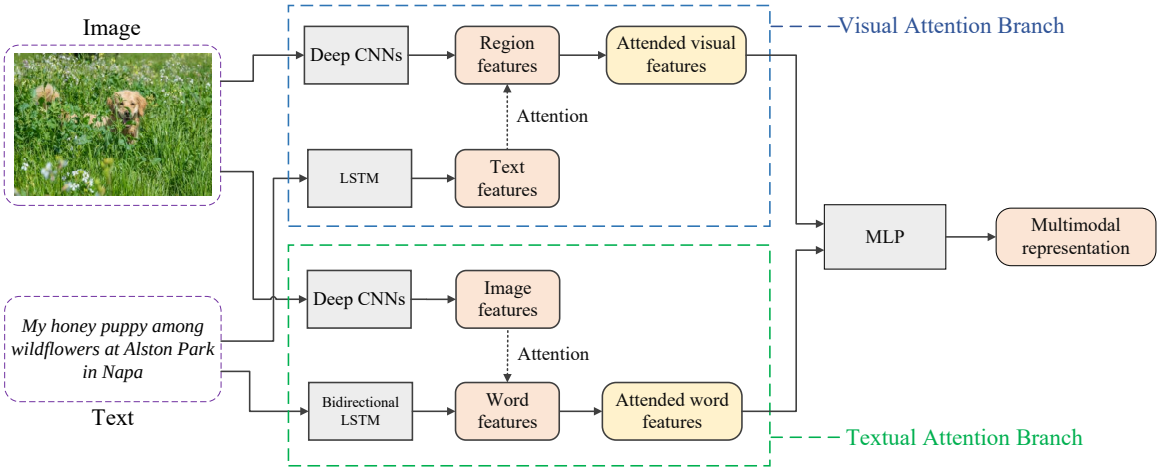


Fig. 3: The neural structure of our two-branch attention model.

where c denotes the dimension of common vector, $W_r \in \mathbb{R}^{c\times e}$ and $W_h \in \mathbb{R}^{c\times h}$ are parameter matrices to be learned, $b_r \in \mathbb{R}^c$ and $b_h \in \mathbb{R}^c$ are bias terms, $\mathbf{1} \in \mathbb{R}^{c\times m\times m}$ is an all 1 matrix that broadcasts the dimension of the left term to $c \times m \times m$, and the symbol $\odot$ indicates an element-wise multiplication or a Hadamard product. The attention scores are then calculated by the convolving operation of the merged feature $F_i^{(c)}$ with the kernel of radius 1 activated by $softmax$ over all the regions as

$$\alpha_i = softmax(W_\alpha * F_i^{(c)} + b_\alpha), \quad \alpha_i \in \mathbb{R}^{m\times m}, \tag{4}$$

where W_α and b_α are trainable convolution parameters, $*$ represents the convolving operation, $softmax$ function is used to normalize attention scores between 0 and 1. In attention map α_i, the attention score of each region is assigned with a value based on the relational degree the corresponding description. Intuitively, we consider that the original region feature should multiply with the attention score at each corresponding visual region. In such way, the importance of each region can be taken into account for feature extraction. Then, the attended region features is computed by the weighted averaging of the original features of the $m \times m$ regions as

$$R_i^{(a)} = \sum_{j=1}^{m\times m} \alpha_{i,j} R_{i,j}, R_i^{(a)} \in \mathbb{R}^e. \tag{5}$$

Compared to the original visual features, the attended visual features have a closer reflection of the correlation to the corresponding description.

2) Textual attention branch: Similarly, we also want to focus on important words related to the corresponding image. For the description T_i, we use pre-trained word embeddings to embed each word and then feed them into a bidirectional LSTM[1] to encode word features as $Q_i \in \mathbb{R}^{q\times l}$, where q is the dimension of word feature and l is the length of the text. As for image V_i, we use pre-trained deep CNNs to extract visual features $P_i \in \mathbb{R}^p$.

[1]We have also tested LSTM but found that Bi-LSTM behaved better.

To correlate the word features Q_i and image features P_i, we first transform them into a common vector space and then fuse them with element-wise multiplication as follows:

$$Q_i' = tanh(W_q Q_i + b_q), \quad Q_i' \in \mathbb{R}^{s\times l}, \tag{6}$$

$$P_i' = tanh(W_p P_i + b_p) \cdot \mathbf{1}, \quad P_i' \in \mathbb{R}^{s\times l}, \tag{7}$$

$$F_i^{(s)} = Q_i' \odot P_i', F_i^{(s)} \in \mathbb{R}^{s\times l}, \tag{8}$$

where s denotes the dimension of common vector, $W_q \in \mathbb{R}^{s\times q}$, $W_p \in \mathbb{R}^{s\times p}$, $b_r \in \mathbb{R}^s$, and $b_h \in \mathbb{R}^s$ are parameters, $\mathbf{1} \in \mathbb{R}^{s\times l}$ is all 1 matrix used to broadcast the dimension of the left term to $s \times l$. Then attention scores are obtained by convolving operation of the merged feature $F_i^{(s)}$ with the kernel of length 1 activated by $softmax$ over all the words, as follows:

$$\beta_i = softmax(W_\beta * F_i^{(s)} + b_\beta), \quad \beta_i \in \mathbb{R}^l, \tag{9}$$

where W_β and b_β are trainable convolution parameters. In attention map β_i, the attention score of each word is assigned with a value based on its relevance to the corresponding image. The attended word features is computed by the weighted averaging of original features of the l words:

$$Q_i^{(a)} = \sum_{j=1}^{l} \beta_{i,j} Q_{i,j}, \quad Q_i^{(a)} \in \mathbb{R}^q, \tag{10}$$

Compared with the original textual features, the attended word features are more effective to reflect the correlation to the corresponding image.

3) Siamese Learning: Through the two attention branches, the attended region features $R_i^{(a)}$ and attended word features $Q_i^{(a)}$ have been obtained. Then, $R_i^{(a)}$ and $Q_i^{(a)}$ are input to a multi-layer perceptron (MLP) to learn the representation of the multimodal contents Y_i.

$$Y_i = mlp(R_i^{(a)}, Q_i^{(a)}), \quad Y_i^{(a)} \in \mathbb{R}^d, \tag{11}$$

where function $mlp(\cdot)$ simulates the neural networks of MLP and d is the dimension of the generated multimodal representation. To make the whole procedure in an end-to-end form,

given the input pair (V_i, T_i), we denote it as a generator function to obtain the representation as

$$Y_i = g(V_i, T_i; \theta_d), \quad Y_i \in \mathbb{R}^d, \tag{12}$$

where θ_d is the parameter of the generator function $g(\cdot)$.

Ideally, we want the two attention networks to assign correct attention scores on visual regions and textual words for multi-modal representation learning, i.e., discover the salient visual region features most related to the corresponding description and focus on important words related to the corresponding image. However, no explicit knowledge is available to learn the alignments. Therefore, to learn the fine-grained correlation between the image-text pair, we need to make the model distinguish the relevance and difference between an image and a text. Motivated by the recent work of visual-textual learning [13], [45], we employ siamese similarity to guide the training of the attention model. For each image, we define the negative text sample as a randomly sampled text which has no relation to the image. For the image and text pair (V_i, T_i), a negative text T_i^- is first sampled. Then both (V_i, T_i) and (V_i, T_i^-) are fed into the attention model. We employ margin ranking loss to learn the matching scores of the positive sample T_i and the negative sample T_i^-.

$$\begin{aligned}&\mathcal{L}_s(\mathcal{V}, \mathcal{T}; \theta_g, \theta_h)\\ &= \sum_{i=1}^{n} max[0, M - h(g(V_i, T_i; \theta_g); \theta_h) + h(g(V_i, T_i^-; \theta_g); \theta_h)],\end{aligned} \tag{13}$$

where function $h(\cdot)$ learns the matching scores given the representation $g(V_i, T_i; \theta_g)$ and $g(V_i, T_i^-; \theta_g)$ generated from the attention branches. We use another MLP activated by $tanh$ to simulate the function $h(\cdot)$. θ_g and θ_h are the parameters shared for both positive and negative samples. This loss function is intended to ensure that the matching score of the positive pair (V_i, T_i) is greater than the negative pair (V_i, T_i^-) such that the fine-grained correlation between the image-text pair can be captured.

For the positive image-text pairs, the image is closely related to its corresponding text descriptions. However, it is usually hard to find the connection between the image and its negative text sample. Then Eq. (4) and Eq. (9) are not appropriate for negative pairs to learn attention weights. Therefore, the calculation of attention weights by Eq. (4) could mislead the visual attention branch to obtain incorrect alignments between visual regions and corresponding text for negative pair (V_i, T_i^-) (similar for Eq. (9)). To tackle this problem, an asymmetrical attending policy is adopted to calculate attention weights:

$$\alpha_i = \begin{cases} softmax(W_\alpha * F_i^{(c)} + b_\alpha) & input : (V_i, T_i) \\ \dfrac{1}{m \times m} & input : (V_i, T_i^-) \end{cases} \tag{14}$$

$$\beta_i = \begin{cases} softmax(W_\beta * F_i^{(s)} + b_\beta) & input : (V_i, T_i) \\ \dfrac{1}{l} & input : (V_i, T_i^-) \end{cases} \tag{15}$$

Through this approach, it can be ensured that negative correspondences are neglected and the attention weights for positive pairs are assigned with reasonable values to find the cross-modal alignments.

C. Adversarial Learning Model

Since the multiple modalities of social images are filled with noisy information as shown in Figure 1, it is necessary to model the uncertainty within social images in the process of representation learning. Generative adversarial networks (GANs) [39] have been proved to be of efficacy in learning representation for image and text [18], [19], [20], [42], [44]. Different from these works, adversarial learning employed in this work act as a regularizer for robust representation learning, which makes the learned representation in accordance with the underlying semantics of the raw data with much noisy information.

GANs are usually composed of two main components: a generator $g(\cdot; \theta_g)$ and a discriminator $d(\cdot; \theta_d)$. In our adversarial learning model, $g(\cdot; \theta_g)$ is the visual-textual attention model which generates representation from multimodal input, while $d(\cdot; \theta_d)$ learns a discriminating function which maps the input following the wanted distribution.

We first designate a prior distribution $p(\mathcal{Z})$ to generate real data, while the generated representation by the attention model is treated as fake samples. During training, the generator is trained to generate representation like prior distribution which is intended to cheat the discriminator, while the discriminator is built to differentiate the generated representation from prior (or true) samples. The generator and discriminator form a two-player game against each other until they reach equilibrium. We formulate the loss function of the discriminator as

$$\begin{aligned}&\mathcal{L}_d(\mathcal{V}, \mathcal{T}, \mathcal{Z}; \theta_d)\\ &= -\sum_{i=1}^{n} log(d(Z_i; \theta_d)) + log(1 - d(g(V_i, T_i; \theta_g); \theta_d)),\end{aligned} \tag{16}$$

where Z_i is a random sample from prior distribution $p(\mathcal{Z})$ and $d(\cdot; \theta_d)$ denotes the possibility of a sample coming from real data. θ_g remains unchanged when discriminator is training. To deceive the discriminator that the generated representation is from prior distribution, the generator is intended to reduce the following loss as

$$\mathcal{L}_g(\mathcal{V}, \mathcal{T}, \mathcal{Z}; \theta_g) = -\sum_{i=1}^{n} log(d(g(V_i, T_i; \theta_g); \theta_d)), \tag{17}$$

where θ_d remains unchanged when the generator is training.

However, the Jenson-Shannon divergence of original GANs easily fall into the instability problem as stated by [22], [46]. To increase training stability, WGAN [22], [40] is proposed with Earth Mover distance (EM distance). EM distance can provide usable and reliable gradient for the loss to achieve synthesis results more easily with better quality. Therefore, we employ WGAN for more stable adversarial learning. based on

original loss of generator (Eq.16) and discriminator (Eq.17), the updated version for WGAN is then written as

$$\mathcal{L}'_d(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_d) = -\sum_{i=1}^{n} d(Z_i;\theta_d) - d(g(V_i,T_i;\theta_g);\theta_d), \tag{18}$$

$$\mathcal{L}'_g(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_g) = -\sum_{i=1}^{n} d(g(V_i,T_i;\theta_g);\theta_d). \tag{19}$$

Note that the choice of the prior distribution $p(\mathcal{Z})$ is also important in adversarial learning. We usually select Gaussian or Uniform as the prior distribution such as [18], [43], [47]. In this work, experiments with two types of prior are conducted, but no significant difference displays. However, different distributions may show big gap of performance on some specific tasks, such as prior domain knowledge-rich tasks.

D. Evolutionary Adversarial Attention Networks

As discussed above, the visual-textual attention model is proposed to excavate the correlation between different modalities and the adversarial learning model is built to regularize the generated representation becoming robust against noise. To make the two models trained into a joint learning procedure, loss Eq.(13) and Eq.(19) are jointly minimized by rewriting the loss function for the generator as

$$\mathcal{L}''_g(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_g,\theta_h) = \mathcal{L}_a(\mathcal{V},\mathcal{T};\theta_g,\theta_h) + \sigma\cdot\mathcal{L}'_g(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_g) + \lambda\cdot\mathcal{L}_{L_2}(\theta_g,\theta_h), \tag{20}$$

where σ and λ are the hyperparameters, and $\mathcal{L}_{L_2}$ is an L2-normalization regularizer to reduce overfitting. Similarly, the loss of discriminator Eq.(16) is rewritten by adding an L2-norm regularizer as

$$\mathcal{L}''_d(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_d) = \sigma\cdot\mathcal{L}'_d(\mathcal{V},\mathcal{T},\mathcal{Z};\theta_d) + \lambda\cdot\mathcal{L}_{L2}(\theta_d). \tag{21}$$

The generator and discriminator can be trained alternatively with stochastic gradient descent (SGD) over the shuffled mini-batches. In the learning step of the generator, the loss function Eq.(20) is minimized to learn parameters θ_g and θ_h. For the discriminator learning step, the samples from the prior distribution and the learned representation generated from the attention model are fed in as input. Then, the parameter set θ_d is updated according to the loss function Eq.(21). We use $gsteps$ and $dsteps$ to represent the number of iterations of generator and discriminator in each training epoch respectively.

For most of the deep learning models, the choice of hyperparameters is conventional done by grid search or manually by the user and often has a significant impact on the performance of the deep learning algorithm. However, grid search of hyperparameters for deep neural networks is very time-consuming, while the efficiency and effectiveness of the manually searching depends heavily on the starting positions across different trials. Inspired by recent work [48], [49] on the evolutionary algorithm on neural networks, we employ evolutionary learning to select the hyper-parameters for our model. The main hyper-parameters include the margin M, balance parameter σ and λ, the number of iterations $gsteps$ and $dsteps$. Each hyperparameter needs to be searched for the model is regarded as a gene for each individual. We also define the range and resolution of each gene to focus on the search space of interest. the population is initialized by sampling the gene values randomly with the uniform distribution. Then the model with each individual or hyperparameter set is trained and the fitness is calculated on the evaluation set. After that, we use the process of selection, recombination, and mutation to generate the next generation according to the fitting degrees of the individuals. Specifically, the selection can directly inherit the optimized individuals and deliver their genes to the next generation. Recombination integrates the parents to form children by simple crossover operations. Mutation is used to insure that there is a diversity among the population. The best individual is selected at the last generations. Basically, the detailed training process of the proposed EAAN is illustrated in Algorithm 1.

Algorithm 1 Training algorithm of EAAN with step size μ, using minibatch SGD for simplicity.

Input: mini-batch Images $\mathcal{V}$,
corresponding mini-batch texts $\mathcal{T}$,
negative texts sampled from $\mathcal{T}$,
prior samples $\mathcal{Z}$,
Output: learned joint multimodal representation generated by the trained model with the best individual.

```
population ← Initialize()
for #generations do
    repeat
        for gsteps do
            θ_g ← θ_g − μ · ∇_{θ_g} L''_g(V, T, Z; θ_g, θ_h)
            θ_h ← θ_h − μ · ∇_{θ_h} L''_g(V, T, Z; θ_g, θ_h)
        end for
        for dsteps do
            θ_d ← θ_d − μ · ∇_{θ_d} L''_d(V, T, Z; θ_d)
        end for
    until model converges
    Evaluate(population)
    parents ← Selection(population)
    offspring ← Recombination(parents)
    population ← Mutation(offspring)
end for
```

IV. Experiments

In this section, the proposed EAAN is evaluated with two tasks, i.e., multi-label classification and tag recommendation, on four real-world datasets.

A. Experimental Preparation

The experiments are performed on the datasets of PASCAL, MIR, CLEF, and NUS-WIDE, which are collected from Flickr and labeled manually. By using the image URLs[2] collected by [50], we crawl the original photos and corresponding descriptions and tags from Flickr. The details are presented as follows:

[2]https://snap.stanford.edu/data/web-flickr.html

- PASCAL [51] is a widely used dataset for the task of image classification. It contains 9,963 images, of which 9,474 images can be found in Flickr.
- MIR [52] contains 1,000,000 images, with only 25,000 annotated. Among them, 13,368 images can be found in Flickr.
- CLEF [53] is a subset of MIR dataset with newly added labels, which contains 18,000 images. Among them, 4,179 of the annotated images can be found in Flickr.
- NUS-WIDE [54] is a benchmark dataset for various vision-related tasks. NUS-WIDE has 269,648 images, of which 226,912 can be found in Flickr.

We remove the images with no groundtruth label or tag. The titles of images are extracted as the textual content. Next, we choose the most frequently appeared 1,000 tags as the tag set for the recommendation experiments. Meanwhile, the image-text pairs containing no words in the tag set are removed. The statistics of these datasets are shown in Table I.

TABLE I: Statistics of the datasets.

	PASCAL	MIR	CLEF	NUS-WIDE
#image	6,151	5,033	3,821	163,862
#label	20	14	99	81
#label per image	1.9	1.92	4.97	2.35
#tag	1,000	1,000	1,000	1,000
#tag per image	4.62	5.98	5.21	9.63

The images are resized to 224×224 with channel RGB as the visual input. Then VGG19 net [55] pre-trained on ImageNet challenge dataset [56] is employed. For visual attention branch, the output features of layer "$conv5\_4$" is used as the region features, of which the dimensionality is $512 \times 14 \times 14$. This means that there are 14×14 regions need to be attended on an image and each region has 512-dimensional features. The output of the last hidden layer of VGG19 is used as the image features for textual attention branch. As to the texts, we employ GloVe [57] to represent each word with a 300-D vector. Both LSTM and Bi-LSTM in the visual-textual attention model are set to have 256 hidden neurons. The MLP to learn multimodal representation is set with the network structure of 1024-1024-512[3] activated with $tanh$. The network structure of MLP in siamese learning is set as $256-128-1$ with $tanh$ activation. For the discriminator of the framework, it is designed as a three-layer MLP, with the structure of $512(tanh)-256(tanh)-2(sigmoid)$. The prior distribution for the adversarial learning is set to be Gaussian distribution with the standard deviation. The hyper-parameters of M, σ, λ, $gsteps$, and $dsteps$ are automatically selected by evolutionary training.

B. Baselines

We compare our models with state-of-the-art methods introduced below:

- **Bimodal-AE** [8]: A bi-modal deep denoising autoencoder to learn multimodal features.
- **M-DBM** [9]: A deep boltzmann machine model to fuse multiple data modalities for representation learning.
- **Corr-AE** [31]: A correspondence autoencoder method to minimize representation learning error and correlation learning error jointly.
- **DCCA** [11]: Deep learning version of CCA which learns nonlinear projection of two views.
- **DSPN** [32]: A deep network combining ranking constraints with structure constraints to learn representation.
- **CMDN** [33]: A two stage multi-view framework which integrates cross-media and intra-media correlation.

Besides, to verify the improvements of our approach compared to our preliminary version AAN [21], we also make a comparison with AAN in the experimental results.

C. Multi-Label Classification

All the datasets used are multi-labeled with unbalanced class distribution. In [58], the metrics for multi-label classification are detailed described. Here we employ macro/micro precision, macro/micro recall, macro/micro F1-measure and mean Average Precision (mAP) as evaluating metrics. To ensure a fair comparison, we randomly split the vectors learned by different methods with the same ratio of 8:1:1 for training, validation, and testing sets respectively. Then a common multi-layer perceptron (MLP) classifier is built for all the methods. we repeat this process 5 times and present average results.

Experimental results on the 4 datasets are illustrated in table II. It shows that the proposed EAAN consistently outperforms other compared methods on all the four datasets. First, from the results of Bimodal-AE and M-DBM, one can see that the two models show relatively worse results due to that they do not fully exploit the nonlinear correlation between different modalities. The results of Corr-AE, DCCA, DSPN, and CMDN indicate that CMDN is the most competitive baseline among them. However, our model EAAN still makes more obvious improvements compared with CMDN. It demonstrates the efficacy of our attention model and adversarial networks on learning multimodal representation. From the comparison of AAN (our preliminary model) and EAAN, one can see that EAAN obtains relatively better metric scores on all the four datasets. It confirms that the two-branch attention network with asymmetrical attending policy proposed in the current version is more effective to excavate the fine-grained correlations between image and text for representation learning. Note that we use Bi-LSTM in the textual attention branch because it shows about 0.1% mAP improvement over LSTM on the datasets. The performance gap is relatively small thus the selection of different LSTMs in the textual attention branch has little effect over the final experimental results.

To further analyze the performance of our model on noisy data, a subset of images is corrupted with additive Gaussian noise of standard deviation of $\sigma = 10$. Then we follow the aforementioned learning and classification procedure to report the results of different multimodal representation learning methods on the dataset NUS-WIDE. The results are shown in Figure 4, which indicate that EAAN performs significantly and consistently better than the baselines. It is worth noting

[3]The number of neurons in the last layer equals to the dimension of the joint multimodal representation d.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEVC.2021.3066285, IEEE Transactions on Evolutionary Computation

TABLE II: Comparison of multi-label classification.

Dataset	Model	Micro-P	Micro-R	Micro-F1	Macro-P	Macro-R	Macro-F1	mAP
PASCAL	Bimodal AE [8]	71.3	34.0	46.0	42.8	39.4	41.0	44.5
	Multimodal DBM [9]	70.9	35.4	47.3	41.5	39.1	40.3	45.2
	Corr-AE [31]	71.2	34.3	46.2	42.6	37.2	39.7	43.3
	DCCA [11]	73.3	37.8	49.9	44.9	42.7	43.8	47.3
	DSPN [32]	74.5	40.0	52.1	45.1	44.3	44.7	48.7
	CMDN [33]	75.0	40.3	52.4	46.1	45.6	45.8	49.4
	AAN [21]	78.9	43.4	56.0	50.1	47.8	49.0	53.4
	EAAN	**81.2**	**44.8**	**57.7**	**52.1**	**48.8**	**50.4**	**54.7**
MIR	Bimodal AE [8]	70.2	69.3	69.7	65.6	65.8	65.7	67.0
	Multimodal DBM [9]	69.7	68.9	69.3	65.7	66.6	66.2	68.6
	Corr-AE [31]	69.5	68.2	68.9	66.9	65.5	66.2	67.3
	DCCA [11]	72.9	72.4	72.6	68.3	66.3	67.3	70.2
	DSPN [32]	70.0	74.0	72.0	69.6	68.5	69.1	71.3
	CMDN [33]	71.0	74.5	72.7	70.2	68.7	69.4	71.8
	AAN [21]	76.9	76.1	76.5	71.5	72.8	72.2	75.3
	EAAN	**79.0**	**77.3**	**78.1**	**72.9**	**74.7**	**73.8**	**77.1**
CLEF	Bimodal AE [8]	51.1	49.3	50.2	40.3	40.0	40.2	34.8
	Multimodal DBM [9]	49.5	48.7	49.1	39.5	40.8	40.1	36.6
	Corr-AE [31]	49.1	48.2	48.6	41.0	39.3	40.2	35.3
	DCCA [11]	53.2	53.1	53.2	42.0	42.7	42.3	40.5
	DSPN [32]	51.7	55.0	53.3	43.6	41.9	42.8	39.3
	CMDN [33]	51.1	55.0	53.0	45.0	42.6	43.8	39.5
	AAN [21]	57.6	55.5	56.5	45.1	46.2	45.7	43.4
	EAAN	**59.5**	**57.8**	**58.7**	**46.5**	**47.9**	**47.2**	**45.1**
NUS-WIDE	Bimodal AE [8]	71.2	51.3	59.7	56.0	49.3	52.5	50.5
	Multimodal DBM [9]	71.0	51.9	59.9	53.6	48.6	51.0	51.2
	Corr-AE [31]	68.3	49.7	57.5	52.0	47.2	49.5	50.3
	DCCA [11]	70.2	50.9	59.0	54.5	48.8	51.5	53.0
	DSPN [32]	72.3	52.5	60.8	58.9	49.9	54.0	55.5
	CMDN [33]	71.7	52.0	60.3	58.1	49.3	53.3	55.4
	AAN [21]	79.0	56.1	65.6	62.4	52.8	57.2	58.3
	EAAN	**81.4**	**57.3**	**67.3**	**64.7**	**54.3**	**59.1**	**59.8**

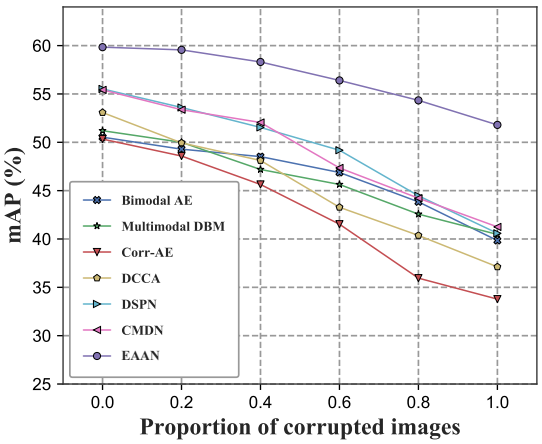


Fig. 4: Multi-label classification results (mAP) on NUS-WIDE with different proportion of corrupted images.

that the margin between EAAN and other methods becomes larger when there are more images corrupted in the dataset. It implies that the proposed method is more robust and stable to learn the joint representation from the multimodal data with much noisy information. It demonstrates the effectiveness of adversarial learning in our model to regularize the representation generating process against uncertainty.

D. Tag Recommendation

We also evaluate EAAN by comparing it with state-of-art methods on the task of tag recommendation, i.e., recommend related tags to a given social image. We filter the most frequent 1,000 tags as the candidate set for recommending. In practice, there are many noisy tags, e.g., some tags have no specific meanings, such as "1" and "funny", some tags have no relation to the social images, some tags have duplicate semantics, such as "cat" and "cats". Thus this task can verify whether the learned representation is effective and robust for tag recommendation.

Table III shows the Micro-F1, Macro-F1, and mAP scores of EAAN and the compared methods on the four datasets. From the results, it can be concluded that the performance of EAAN is better than Bimodal AE, M-DBM, Corr-AE, DCCA, DSPN, and CMDN on all metrics. Since DSPN and CMDN are state-of-the-art multimodal embedding methods, it validates the effectiveness of our adversarial visual-textual attention model for joint multimodal representation learning. On the other side, these four datasets have different number of data and instances, thus the improvement of our approach over compared methods shows the generality of EAAN.

E. Visualization of learned attentions

One advantage of including the attention mechanism is the ability to visualize what the model "sees". To better understand the interpretability of meaningful attention drawn

TABLE III: Tag recommendation results on four datasets.

Dataset	Model	Micro-F1	Macro-F1	mAP
PASCAL	Bimodal AE [8]	20.4	17.8	19.1
	Multimodal DBM [9]	18.4	17.1	18.3
	Corr-AE [31]	19.9	18.7	19.4
	DCCA [11]	21.5	20.0	21.0
	DSPN [32]	24.2	21.0	22.4
	CMDN [33]	22.8	21.3	22.7
	AAN [21]	25.1	22.6	24.3
	EAAN	**28.7**	**25.1**	**26.5**
MIR	Bimodal AE [8]	17.2	15.4	16.4
	Multimodal DBM [9]	19.2	16.1	17.9
	Corr-AE [31]	16.8	14.8	16.1
	DCCA [11]	20.1	17.5	18.8
	DSPN [32]	20.2	18.0	19.5
	CMDN [33]	21.0	18.2	19.4
	AAN [21]	24.2	20.0	22.2
	EAAN	**24.3**	**21.8**	**23.9**
CLEF	Bimodal AE [8]	9.0	6.9	8.2
	Multimodal DBM [9]	10.5	8.0	9.2
	Corr-AE [31]	9.8	7.7	8.9
	DCCA [11]	12.2	8.7	10.8
	DSPN [32]	13.8	9.5	12.4
	CMDN [33]	13.1	10.9	12.6
	AAN [21]	14.7	13.2	14.1
	EAAN	**15.8**	**15.1**	**15.3**
NUS-WIDE	Bimodal AE [8]	23.1	20.9	21.3
	Multimodal DBM [9]	23.4	22.0	23.2
	Corr-AE [31]	23.0	21.9	22.1
	DCCA [11]	25.1	21.6	24.3
	DSPN [32]	27.1	22.5	24.8
	CMDN [33]	26.6	23.2	25.0
	AAN [21]	28.8	25.3	28.4
	EAAN	**31.7**	**28.4**	**30.5**

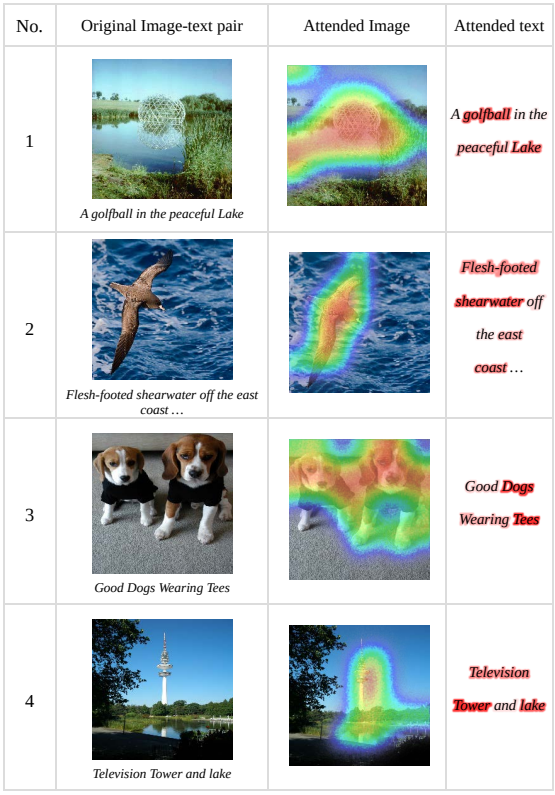


Fig. 5: Examples of learned attentions. For each sample, we present the original image-text pair, the attended image and text. Red areas indicate the attended regions and red-stroking are drawn on attended words.

from the proposed two-branch attention model, we present four examples in Figure 5. For the visual attention, similar to [14], we first up-sample the attention scores and then use a Gaussian filter. Different from [14], the up-sampled attention scores are further drawn with a heat map for more colorful visualization. The final images drawn with attention are the original images masked by the heat map with the transparency of 0.7. For the textual attention, we simply red-stroke each word with the corresponding attention scores. Therefore, if the attention scores of the words are greater, the words are colored redder.

From the figure, it can be seen that our model draws the right attention to the images and texts. For the first example, the sentence "A golfball in the peaceful Lake" is drawn with greater attentions on the words "golfball" and "Lake" while the image is also drawn on the more importance regions which reflect the semantic information of the text. The proposed attention mechanism makes the relation between multimodal contents explicit and interpretable, and the fine-grained correlation between the multimodal contents is encoded to obtain a more effective representation.

F. Parameter Sensitivity

In this subsection, we test the parameter sensitivity of EAAN and present the results of mAP on the task of multi-label classification with different parameter settings on NUS-WIDE. Specifically, we evaluate how different balance parameters (σ and λ) and embedding dimensions (d) influence the experimental results.

Embedding dimension: We first fix $\sigma = 0.3$ and $\lambda = 0.01$ and change the dimensionality of the representation. From Figure 6, it can be seen that better performance can be achieved at the beginning with a bigger dimension. This is reasonable because more information can be encoded with more bits. However, as the embedding dimension continuously boosts, the performance of the model starts to deteriorate. This is because that a too large dimension could also introduce noisy information. Overall, it is important to choose an appropriate embedding dimension size, and EAAN reaches the best mAP score at about $d = 256$.

Balance parameters: In the model, we use σ to regulate the importance of adversarial learning and λ as a trade-off for L2-norm regularizer. We fix the dimensionality $d = 256$ and test the performance with different σ and λ. Based on the curves in Figure 7, one can see that setting a trade-off term is needed to reduce overfitting because the model performs relatively poorly when $\lambda = 0$. However, too big L2-norm term also affects the process of representation learning. For another, the regulating parameter σ shows a more significant influence on the performance. When $\sigma = 0$, only the visual-textual attention model is optimized, which means the model is not capable to learn from adversarial networks. When σ becomes larger, the model concentrates more on the optimization of adversarial learning. From the curve, it can be seen that the best performance is obtained at $\sigma = 0.3$ and $\lambda = 0.01$.

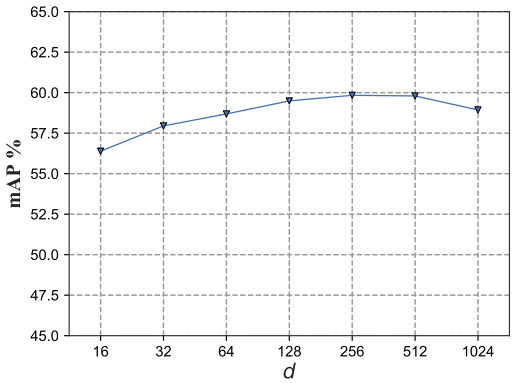


Fig. 6: Parameter sensitivity study for the embedding dimension.

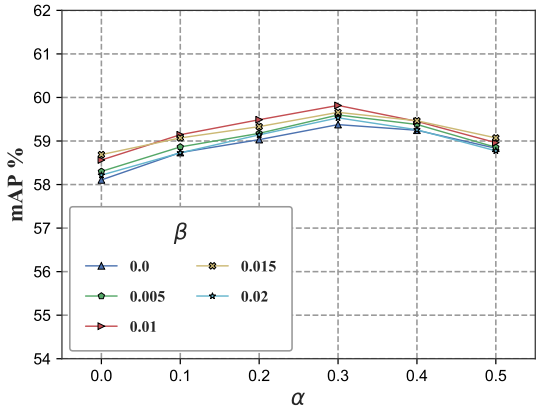


Fig. 7: Parameter sensitivity study for the balance parameters of α and β.

G. Ablation Study

To analyze the effectiveness of the components used in the proposed model for addressing noise, we ablate the proposed model and conduct multi-label classification experiments on both original data and corrupted data. Specifically, we re-train our method by ablating different components: 1) visual-textual attention networks (VTA), traditional generative adversarial networks (GAN), and Wasserstein generative adversarial networks (WGAN).

TABLE IV: The ablative results (mAP) on the original NUS-WIDE and corrupted NUS-WIDE.

Ablation Model	NUS-WIDE	Corrupted NUS-WIDE
SM	53.1	47.7
VTA	59.4	54.2
VTA+GAN	59.6	56.6
VTA+WGAN	**59.8**	**57.3**

The ablation results (mAP) for the task of multi-label classification on the original NUS-WIDE and corrupted NUS-WIDE are shown in Table IV. The simple model (SM) is a simple version of our method by changing the two kinds of attention mechanism with two average pooling layers. From the results, one can see that VTA outperforms SM by over 6% on the metric of mAP. It validates that the visual-textual attention model is useful to exploit the correlation for joint representation learning. Both VTA+GAN and VTA+WGAN outperforms VTA by employing the generative adversarial networks to regularize the representation for uncertainty learning. Especially, one can see that the adversarial learning behaves more effective on the corrupted NUS-WIDE than original dataset. By comparing the two types of GANs, VTA+WGAN shows slight improvements over VTA+GAN. The reason is that WGAN supplies for more stable and effective training.

V. Conclusions and Future Work

In this paper, we propose a multimodal representation learning method named Evolutionary Adversarial Attention Networks. It combines the attention mechanism with the adversarial networks through evolutionary training to learn the representation more effectively and robustly. Specifically, a two-branch visual-textual attention model with siamese learning is proposed to exploit the fine-grained correlation between different modalities. Then the adversarial learning model is employed to regularize the representation generated by the attention model. Next, the two models are optimized jointly in a holistic evolutionary learning framework to learn the representation. We evaluate our approach on four real-world datasets with two tasks. The results demonstrate the efficacy of the proposed EAAN on learning robust representation for multimodal data.

In the future, we want to generalize our method to other types of multimodal data such as voice and videos. Besides, we will explore how to fuse other information, e.g., the social links among images and the relationship among image owners, for more effective learning for multimodal representation.

References

[1] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, A. Hanjalic, C. Snoek, M. Worring, D. C. A. Bulterman, B. Huet, A. Kelliher, Y. Kompatsiaris, and J. Li, Eds. ACM, 2016, pp. 978–987.

[2] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Evolving deep convolutional neural networks for image classification," *IEEE Trans. Evol. Comput.*, vol. 24, no. 2, pp. 394–407, 2020.

[3] Y. Zhang, J. Wu, Z. Cai, and S. Y. Philip, "Multi-view multi-label learning with sparse feature selection for image annotation," *IEEE Transactions on Multimedia*, 2020.

[4] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, J. Huai, R. Chen, H. Hon, Y. Liu, W. Ma, A. Tomkins, and X. Zhang, Eds. ACM, 2008, pp. 327–336.

[5] Y. S. Rawat and M. S. Kankanhalli, "Contagnet: Exploiting user context for image tag recommendation," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, A. Hanjalic, C. Snoek, M. Worring, D. C. A. Bulterman, B. Huet, A. Kelliher, Y. Kompatsiaris, and J. Li, Eds. ACM, 2016, pp. 1102–1106.

[6] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 21–29.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 289–297.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, L. Getoor and T. Scheffer, Eds., 2011, pp. 689–696.

[9] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 2231–2239.

[10] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 1247–1255.

[11] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3441–3450.

[12] J. Weston, S. Bengio, and N. Usunier, "WSABIE: scaling up to large vocabulary image annotation," in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, T. Walsh, Ed. IJCAI/AAAI, 2011, pp. 2764–2770.

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 2121–2129.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 2048–2057.

[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6077–6086.

[16] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1979–1988.

[17] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds. ACM, 2018, pp. 1398–1406.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[19] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1283–1292.

[20] J. Glover, "Modeling documents with generative adversarial networks," *CoRR*, vol. abs/1612.09122, 2016. [Online]. Available: http://arxiv.org/abs/1612.09122

[21] F. Huang, X. Zhang, and Z. Li, "Learning joint multimodal representation with adversarial attention networks," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds. ACM, 2018, pp. 1874–1882.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 214–223.

[23] T. Bai, Y. Li, and X. Zhou, "Learning local appearances with sparse representation for robust and fast visual tracking," *IEEE Trans. Cybernetics*, vol. 45, no. 4, pp. 663–675, 2015.

[24] D. G. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist, "Conceptvector: Text visual analytics via interactive lexicon building using word embedding," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 361–370, 2018.

[25] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. IEEE Computer Society, 2010, pp. 902–909.

[26] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4651–4659.

[27] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, and A. F. Smeaton, Eds. ACM, 2003, pp. 127–134.

[28] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*. AUAI Press, 2005, pp. 633–641.

[29] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 251–260.

[30] H. Suk, S. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[31] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. Natsev, and W. Zhu, Eds. ACM, 2014, pp. 7–16.

[32] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5005–5013.

[33] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, S. Kambhampati, Ed. IJCAI/AAAI Press, 2016, pp. 3846–3853.

[34] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, Q. Liu, R. Lienhart, H. Wang, S. K. Chen, S. Boll, Y. P. Chen, G. Friedland, J. Li, and S. Yan, Eds. ACM, 2017, pp. 154–162.

[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[36] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1243–1252.

[37] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds. The Association for Computational Linguistics, 2015, pp. 1412–1421.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6450–6458.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in

Advances in neural information processing systems, 2014, pp. 2672–2680.

[40] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5769–5779.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[42] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[43] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015. [Online]. Available: http://arxiv.org/abs/1511.05644

[44] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[45] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014. [Online]. Available: http://arxiv.org/abs/1411.2539

[46] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[47] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *CoRR*, vol. abs/1605.09782, 2016. [Online]. Available: http://arxiv.org/abs/1605.09782

[48] Y. Sun, G. G. Yen, and Z. Yi, "Evolving unsupervised deep neural networks for learning meaningful representations," *IEEE Trans. Evol. Comput.*, vol. 23, no. 1, pp. 89–103, 2019.

[49] Y. Li, Z. Zhan, Y. Gong, W. Chen, J. Zhang, and Y. Li, "Differential evolution with an evolution path: A DEEP evolutionary algorithm," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1798–1810, 2015.

[50] J. J. McAuley and J. Leskovec, "Image labeling on a network: Using social-network metadata for image classification," in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7575. Springer, 2012, pp. 828–841.

[51] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[52] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 39–43.

[53] S. Nowak and M. J. Huiskes, "New strategies for image annotation: Overview of the photo annotation task at imageclef 2010," in *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, ser. CEUR Workshop Proceedings, M. Braschler, D. Harman, and E. Pianta, Eds., vol. 1176. CEUR-WS.org, 2010.

[54] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from national university of singapore," in *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*, S. Marchand-Maillet and Y. Kompatsiaris, Eds. ACM, 2009.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[56] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255.

[57] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1532–1543.

[58] X. Wu and Z. Zhou, "A unified view of multi-label performance measures," *CoRR*, vol. abs/1609.00288, 2016.

Feiran Huang received his B.Sc. degree from Central South University, Changsha, China in 2011. He received his Ph.D. degree in computer software and theory from School of Computer Science and Engineering, Beihang University, Beijing, China, in 2019. He is currently an assistant professor at School of Information Science and Technology & College of Cyber Security, Jinan University, Guangzhou, China. He has published over 30 papers in international conferences and journals, such as IEEE TIP, IEEE TCYB, IEEE T-ITS, IEEE TNNLS, ACM TOMM, ACM MM, and ACM ICMR. His research interests include social media analysis and multi-modal learning. He is a member of IEEE and ACM.

Alireza Jolfaei received the Ph.D. degree in Applied Cryptography from Griffith University, Gold Coast, Australia. He is the Program Leader of Master of IT in Cyber Security at Macquarie University, Sydney, Australia. His main research interests are in Cyber and Cyber-Physical Systems Security. He has participated in several projects involving different aspects of Cyber Security. On these topics he has published over 80 papers appeared in journals, conference proceedings, and books. Before Macquarie University, he has been a faculty member with Federation University Australia and Temple University in Philadelphia, PA, USA. He has received multiple awards for Academic Excellence, University Contribution, and Inclusion and Diversity Support. He received the prestigious IEEE Australian council award for his research paper published in the IEEE Transactions on Information Forensics and Security. He served as the Chairman of the Computational Intelligence Society in the IEEE Victoria Section and also as the Chairman of Professional and Career Activities for the IEEE Queensland Section. He has served as the guest associate editor of IEEE journals and transactions, including the IEEE IoT Journal and IEEE Transactions on Intelligent Transportation Systems. He has served as a program co-Chair and a Technical Program Committee member, for major conferences in Cyber Security. He is a Distinguished Speaker of ACM on the topic of Cyber Security and a Senior Member of IEEE.

Ali Kashif Bashir is a Senior Lecturer/Associate Professor and Program Leader of BSc (H) Computer Forensics and Security at the Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom. He is also with School of Electrical Engineering and Computer Science, National University of Science and Technology, Islamabad (NUST) as an Adjunct Professor and School of Information and Communication Engineering, University of Electronics Science and Technology of China (UESTC) as an Affiliated Professor and Chief Advisor of Visual Intelligence Research Center, UESTC. He is a senior member of IEEE, member of IEEE Industrial Electronic Society, member of ACM, and Distinguished Speaker of ACM. His past assignments include Associate Professor of ICT, University of the Faroe Islands, Denmark; Osaka University, Japan; Nara National College of Technology, Japan; the National Fusion Research Institute, South Korea; Southern Power Company Ltd., South Korea, and the Seoul Metropolitan Government, South Korea. He received his Ph.D. in computer science and engineering from Korea University South Korea. He has authored over 180 research articles; received funding as PI and Co-PI from research bodies of South Korea, Japan, EU, UK and Middle East; supervising/co-supervising several graduate (MS and PhD) students. His research interests include internet of things, wireless networks, distributed systems, network/cyber security, machine learning, etc. He is serving as the Editor-in-chief of the IEEE FUTURE DIRECTIONS NEWSLETTER. He is also serving as area editor of KSII Transactions on Internet and Information Systems; associate editor of IEEE Internet of Things Magazine, IET Quantum Computing. He is leading many conferences as a chair (program, publicity, and track) and had organized workshops in flagship conferences like IEEE Infocom, IEEE Globecom, IEEE Mobicom, etc.