Bi-Directional Spatial-Semantic Attention Networks for Image-Text Matching

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, and Zhoujun Li

***Abstract*— Image-text matching by deep models has recently made remarkable achievements in many tasks, such as image caption and image search. A major challenge of matching the image and text lies in that they usually have complicated underlying relations between them and simply modeling the relations may lead to suboptimal performance. In this paper, we develop a novel approach bi-directional spatial-semantic attention network, which leverages both the word to regions (W2R) relation and visual object to words (O2W) relation in a holistic deep framework for more effectively matching. Specifically, to effectively encode the W2R relation, we adopt LSTM with bilinear attention function to infer the image regions which are more related to the particular words, which is referred as the W2R attention networks. On the other side, the O2W attention networks are proposed to discover the semantically close words for each visual object in the image, i.e., the visual O2W relation. Then, a deep model unifying both of the two directional attention networks into a holistic learning framework is proposed to learn the matching scores of image and text pairs. Compared to the existing image-text matching methods, our approach achieves state-of-the-art performance on the datasets of Flickr30K and MSCOCO.**

***Index Terms*— Image-text matching, attention networks, deep learning, spatial-semantic.**

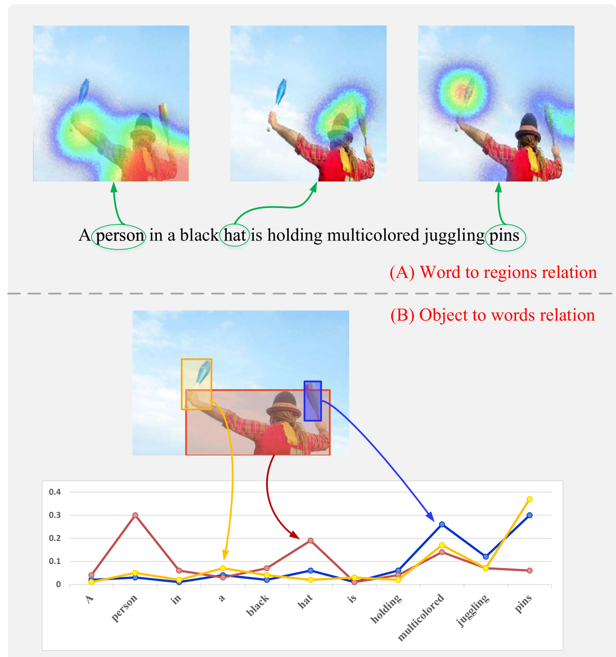

Fig. 1. An illustration of bi-directional relations between image and text on the dataset Flick30K. (A) Word to regions relation denotes what visual regions a word is more related to. (B) Object to words relation denotes what words a visual object is more related to.

I. Introduction

With the rapid growth of natural language processing (NLP) and computer vision (CV) technology, the machine is expected to understand the semantics of languages and images in the near future. Automatically describing the visual data with natural language is useful for many tasks, such as image annotation and captioning [1]–[6], and a more natural way for image search by retrieving images with a sentence or phrase as query [7]–[10]. The association between images and textual descriptions can be formalized as an image-text matching problem, i.e., the semantically related image-text pairs should have higher matching scores than those unrelated pairs.

Due to the heterogeneous representations and the complicated relations between the visual content and text description, image-text matching remains challenging. To address the first problem, many methods based on deep models (e.g., CNN, RNN) are proposed to project the heterogeneous representations into a shared embedding space for similarity comparison or fuse the two features to learn the matching scores [11]–[14]. However, how to effectively model the complicated relations is still an open problem. There exist bi-directional and fine-granularity relations between image and text. Usually, each word is related to some specific regions of the images, and each visual object is also related to some specific words in the text content. Figure 1 shows an example of a sentence "A person in a black hat is holding multicolored juggling pins" and the corresponding image from the Flick30K dataset.

Manuscript received February 1, 2018; revised July 4, 2018; accepted November 11, 2018. Date of publication November 19, 2018; date of current version December 12, 2018. This work was supported in part by the Beijing Natural Science Foundation of China under Grant 4182037, in part by the National Natural Science Foundation of China under Grants U1636210 and U1636211, and in part by the Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2017ZX-19. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aydin Alatan. *(Corresponding author: Xiaoming Zhang.)*

F. Huang is with the Beijing Key Laboratory of Network Technology, Beihang University, Beijing 100191, China.

X. Zhang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China (e-mail: yolixs@buaa.edu.cn).

Z. Zhao is with the National Computer Emergency Technical Team/Coordination Center of China, Beijing 100029, China.

Z. Li is with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.

Digital Object Identifier 10.1109/TIP.2018.2882225

For the word “person”, “hat”, and “pins”, there is a small part of visual regions reflecting their semantics in the top 3 images of Figure 1. On the other side, for the objects enclosed by the yellow box, blue box and red box in Figure 1, only several words are more related respectively. Therefore, if the two types of relations are effectively exploited to highlight the correspondence between visual regions and words, an image can be matched with a more related text and vice versa.

There have been many methods on image-text matching, which can be divided into the following two categories. The first category is Canonical Correlation Analysis (CCA)-based methods [15], which aims to find a linear projection that maximizes the correlation between the projected vectors from image and text. A three-view embedding approach is proposed in [16] to fuse visual content, tags, and semantic information for cross-modal matching and image-to-image matching. However, it is difficult for these methods to effectively capture the non-linear relation between image and text. DCCA proposed in [17] and [18] extends the CCA with a deep learning model. However, as pointed out in [19], DCCA is hard to scale up to large datasets due to the unstable training for the generalized eigenvalue problem. The second type is ranking-based methods, which aims to learn a ranking loss function from image and text pairs. WSABIE [20] and DeVISE [21] learn a linear transformation to project the image and text features to a shared embedding space with a ranking loss function. The work in [22] proposes multimodal CNNs (m-CNNs) to match image and sentence at different semantic levels. However, these methods cannot well exploit the fine-grained correlation between visual regions and text words.

On the other side, attention mechanism makes an alignment between the input and output [3], which can highlight the remarkable features as needed. Attention mechanism has been studied extensively in both vision and language problems including image caption generation [3]–[5], [23], VQA [24]–[27], image classification [28]–[30], and machine translation [31], [32]. The visual attention mechanism adaptively focuses on specific discriminative local regions rather than spreads evenly over the whole image [3], [11]. The textual attention mechanism adaptively selects important words [11], [31]. Attention mechanism has made significant improvements on a wide range of applications through deep architectures such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Recently, Dual Attention Networks [11] is proposed to learn two representations for image and text separately. However, the attention networks are built in two branches for two separate modalities, which neglects the cross-modal correlation between image and text.

In this paper, we take full advantage of the attention mechanism to effectively explore the fine-granularity relations between image and text for image-text matching. In particular, we investigate: (1) how to effectively encode the bi-directional relations between image and text? (2) how to effectively exploit the fine-granularity correlation between the multimodal contents for image-text matching? To solve these questions, we propose a novel image-text matching approach named Bi-directional Spatial-Semantic Attention Networks (BSSAN), in which the salient features of both the visual content and textual description are highlighted to improve image-text matching. Specifically, a spatial-semantic attention model is proposed to encode the bi-directional correlation between image and text. It contains two attention networks, i.e., the word to regions (W2R) attention network and the object to words (O2W) attention network. W2R is a spatial-based attention module which learns the salient image regions for the corresponding text words, while O2W is a semantic-based attention module which attends to discover the semantically related words for the image objects. Then the two directional attention networks are unified into a holistic learning framework. A matching score is learned with the deep model by exploiting the two-directional relations for a pair of image and text. The main contributions are summarized as follows:

- We investigate the problem of matching the image and text by exploiting the bi-directional and fine-granularity correlations between visual regions and textual words.
- We propose a novel deep model for image-text matching, i.e., Bi-directional Spatial-Semantic Attention Networks (BSSAN). Our model naturally combines both the word to regions (W2R) relation and object to words (O2W) relation for image-text matching score learning.
- The proposed model is extensively evaluated on the datasets of Flickr30K and MSCOCO with the cross-modal search task. The experimental results confirm the superiority of BSSAN against state-of-the-art baselines.

The remainder of this paper is organized as follows. Section 2 reviews the related works, and Section 3 introduces the approach BSSAN in details. Then, the experimental results are presented in Section 4. Finally, we make conclusions in Section 5.

II. Related Work

In the computer vision and multimedia communities, jointly modeling visual content and text has been an active research topic. It supports various applications, such as image caption [1]–[5], [23] and visual question answering [24]–[27], [33], [34]. In this section, we first introduce the general attention mechanism on both visual and textual researches. Then, we summarize the most recent advances of image-text matching from two dimensions.

A. Attention Mechanisms

Attention mechanism has made significant progress in the field of computer vision [3]–[5], [35] and natural language processing [31], [32], [36]. In machine translation, BRNN [31] is proposed to align a source sentence with the corresponding target sentence. BRNN can automatically predict target words by searching the most related parts of a source sentence. In image captioning, Xu *et al.* [3] exploit two attention mechanisms, i.e., soft-attention and hard-attention, to learn to describe the content of images. In [4], attention is targeted on a set of concepts extracted from the image. A set of top-down visual features are used to guide when and where the attention should be drawn to generate image captions. In visual question answering, several models [11], [24]–[27] have been proposed to assign attention on image regions or textual questions when generating an answer. Stacked attention networks (SANs) [27] uses a multiple-layer attention mechanism to infer the answer progressively by querying the image multiple times. Lu *et al*. [25] exploit the co-attention strategy to assign attention to jointly reasons about image and question attention. Recently, Dual Attention Networks [11] is proposed for image-text matching, which refines the visual and textual attention via multiple reasoning steps. Our work is different from the dual attention model which is built on two uni-modal networks separately. The attention model is built with the bi-directional networks in our approach, which can well exploit the cross-modal correlation and reinforce the salient features of the two modalities complementally.

B. Image-Text Matching

1) CCA-Based Methods: Methods based on Canonical Correlation Analysis (CCA) have been successfully applied in image-text matching [15], [37], [38]. It projects the features of two views into a shared vector space by maximizing the cross relation between them. The kernel generalization of CCA named KCCA [15] has been proposed to find non-linear relations between datasets, in which kernel methods are used implicitly to perform the non-linear transformation. Klein *et al*. [39] show that properly normalized CCA can achieve satisfying performance by utilizing state-of-the-art visual and textual features. The main drawbacks of CCA are its heavy memory consumption and ineffectiveness to capture the non-linear relation between image and text. Compared to hand-crafted objectives, Deep Canonical Correlation Analysis (DCCA) [13], [14], [17], [18] optimizes the objective with the deep learning method, in which the entire correlation can be maximized through optimizing the matrix trace norm. This allows an end-to-end learning to propagate the gradient down in a deep learning framework. However, as pointed out in [19], SGD cannot well solve the generalized eigenvalue problem due to the unstable covariance estimation.

2) Ranking-Based Methods: The second type of methods learns a joint representation or a matching score on the neural networks with a ranking loss. WSABIE [20] and DeVISE [21] utilize a single-directional ranking loss to learn the linear projections of image and text features to the common subspace. In [40], an image-text embedding method is proposed to train a two-branch deep network with a margin-based objective function. The object function consists of the structure-preserving terms and the bi-directional rank terms. Ma *et al*. [22] combine images and sentence fragments into a joint representation to infer the matching scores with a CNN model. Karpathy *et al*. [41] explicitly compute all pairwise distances to automatically calculate the alignments between visual regions and textual fragments. CMDN [42] 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. Though these methods have achieved encouraging performance on image-text matching, they cannot well exploit the fine-granularity and bi-directional correlation between visual regions and text words.

III. Problem Statement

Before the problem formulation, we define several notations used in the paper. There are two modalities of data considered in this paper, namely images and text descriptions. Let $\mathcal{V} = \{V_1, \ldots, V_i, \ldots, V_n\}$ and $\mathcal{T} = \{T_1, \ldots, T_i, \ldots, T_n\}$ denote n samples of images and text documents respectively. Then the target is to learn a matching scoring function, such that for each image V_i, the pair of the image and the matched text (V_i, T_i) should have a higher matching score than the pair of the image and an unmatched text (V_i, T_i^-).

The framework of BSSAN is illustrated in Figure 2. BSSAN consists of three major components. First, the word to regions (W2R) attention network is proposed to infer the image regions which is attended by the corresponding words. Local image region representations are first extracted from the pre-trained convolutional layers and repeatedly fed into the LSTM together with each word by the attention module. Attention score of each region is obtained by a bilinear function which enables the spatial information of a region to be encoded from surrounding context. Second, the object to words (O2W) attention network aims to discover the semantically-close words for the objects in the image. Specifically, we generate object proposals using Faster-RCNN [43] and extract high-level features of each object from the pre-trained CNNs. Attention scores for each object are learned through the multi-layer perceptron. These attention scores reflect the underlying mapping between a given image object and the related linguistic concepts. Then a deep model unifying the two directional attention networks into a holistic learning framework is proposed to learn an integrated matching score of the image and text pair.

IV. Bi-Directional Spatial-Semantic Attention Networks

In this section, we first introduce how to encode the bi-directional relations between image and text using W2R and O2W attention networks. Then we discuss how to integrate the two attention networks with a deep model to learn the matching scores for image and text pairs.

A. Word to Regions (W2R) Attention Network

Usually, for each word, there is a part of visual regions that are more related to the semantics of this word. If the features of these visual regions are highlighted, the matching score of the text and the related image could be enhanced. Therefore, W2R attention network is proposed to infer the visual attention on image regions attended by the corresponding word. Due to the success of fine-grained visual representation and visualization, visual attention has gained great benefits in many visual tasks, such as image classification [28], [30] and image caption [3]. In contrast to these works, our attention model is formed in a multimodal structure. We use attention-based LSTM to excavate the attention relations between visual regions and sequential words, which can "stick out" the related part of the image regions for each word to support image-text matching.

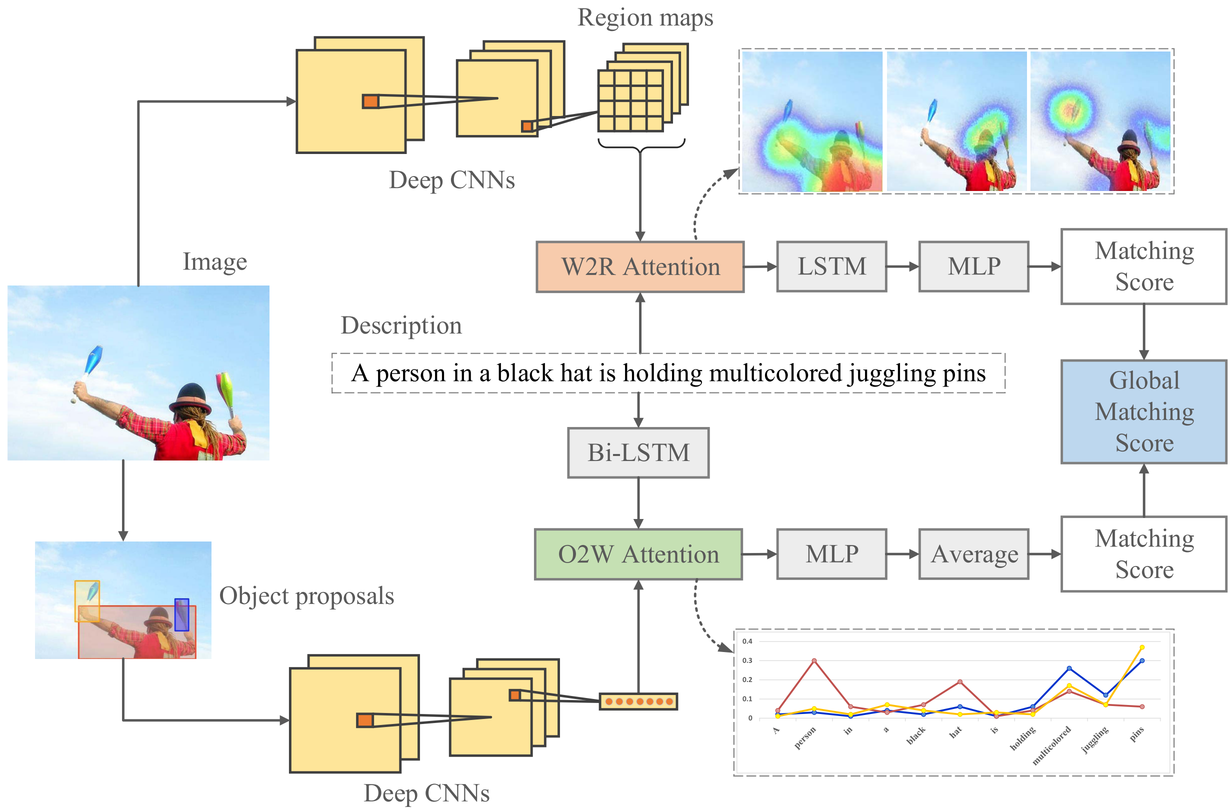

Fig. 2. The framework of BSSAN for image-text matching.

Let $T_i = \{t_i^1, \ldots, t_i^k, \ldots, t_i^L\}^T, T_i \in \mathbb{R}^{L\times E}$ denotes the text features of the i-th image-text pair (V_i, T_i), which is a sentence of length L, where each word is a pre-trained embedding vector. k denotes the index of a word and E is the dimension of the word embedding. For the image V_i, the pre-trained deep CNNs are employed to acquire the visual region maps $R_i = \{r_{i,1}, \ldots, r_{i,j}, \ldots, r_{i,D}\} \in \mathbb{R}^{D\times M}$ where M is the map dimension and D is the number of regions. For each word t_i^k, the W2R attention network assigns a score $\alpha_{i,j}^k \in [0, 1]$ to each visual region $r_{i,j}$ based on the extent it relates to word t_i^k. A $softmax$ function is used to calculate $\alpha_{i,j}^k$ as follows:

$$\alpha_{i,j}^k = \frac{\exp(e_{i,j}^k)}{\sum_{j=1}^{D} \exp(e_{i,j}^k)} \tag{1}$$

where

$$e_{i,j}^k = \varphi((t_i^k)^T W r_{i,j} + b) \tag{2}$$

is the unnormalized attention score measuring the relevance of the visual region $\{r_j\}$ and the word t_i^k. W is the weight matrix and b is the bias term, which are the parameters to be learned. $\varphi(\cdot)$ is the $tanh$ activation function to capture the non-linear correlation. $\alpha_{i,\cdot}^k$s normalize the attentions of all the regions $\{r_{i,j}\}_{1\leqslant j\leqslant D}$ drawn from word t_i^k.

Then we use $\alpha_{i,\cdot}^k$s to adjust the attentive strength over different image regions. By weighted summing the image regions for word t_i^k, the attended visual features can be calculated as follows:

$$u_i^k = \sum_{j=1}^{D} \alpha_{i,j}^k r_{i,j}, \quad U_i \in \mathbb{R}^{L\times M} \tag{3}$$

The spatial attention process to automatically generate the attended image features is formulated as follows:

$$U_i = f_a(T_i, R_i; \theta_a), \quad U_i \in \mathbb{R}^{L\times M} \tag{4}$$

where θ_a is the weight parameters which consists of W and b in Eq.(2). In contrast to the original visual features which are shared by all words, u_i^k is more representative to show word-related region features. We make concatenation of u_i^k and t_i^k and denote it as $c_i^k = [u_i^k; t_i^k]$, which is then input to the k-th cell of the LSTM. With this method, visual and textual features can be integrated as a joint sequence $\{c_i^1, \ldots, c_i^k, \ldots, c_i^L\}$ fed into the LSTM network. Then we further build a multi-layer perceptron (MLP) after the last LSTM cell to learn the matching score of the word to regions (W2R) network. The architecture of the W2R attention network is illustrated in Figure 3. Let $Y_i \in \mathbb{R}^1$ be the matching score:

$$Y_i = f_l(T_i, U_i; \theta_l), \quad Y_i \in \mathbb{R}^1 \tag{5}$$

where f_l is the function to integrate LSTM and MLP and θ_l is the weight parameters.

To obtain an end-to-end process, we pipeline the two functions mentioned in Eq.(4) and Eq.(5) to obtain a unique function. The function maps the multimodal input to a matching score Y_i directly as follows:

$$Y_i = f_{w2r}(V_i, T_i; \theta_{w2r}) \tag{6}$$

where f_{w2r} is the pipeline of f_a and f_l, and θ_{w2r} is the parameter set of θ_a and θ_l.

We hope that the correct attention weights can be assigned to the related image regions for each word. However, no

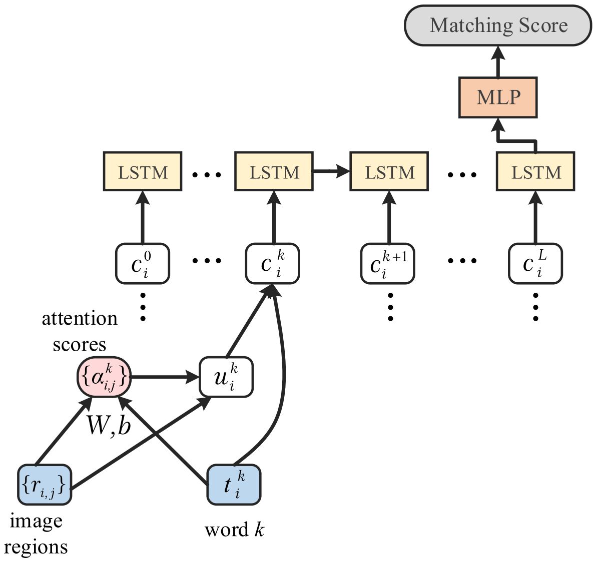

Fig. 3. The architecture of W2R attention network. The attention process produces $\{c_i^1, \ldots, c_i^k, \ldots, c_i^L\}$, where each c_i^k is the concatenation of the attended image features u_i^k and the corresponding word vector t_i^k. They are fed into LSTM cells, and MLP is used to produce the final matching score.

explicit knowledge can be found to guide the alignments. To handle this problem, we employ siamese network structure [21], [44] to learn the W2R attention networks. Given an image V_i and corresponding text T_i, we first sample an unmatched description as the negative text T_i^-. Then positive pair (V_i, T_i) and negative pair (V_i, T_i^-) are input to the W2R attention network to learn the positive matching scores Y_i and the negative matching score Y_i^- respectively. The pairwise ranking loss can be formulated as follows:

$$\begin{aligned} L_1(\mathcal{V}, \mathcal{T}; \theta_{w2r}) &= \sum_{i=1}^{n} max[0, M - Y_i + Y_i^-] \\ &= \sum_{i=1}^{n} max[0, M - f_{w2r}(V_i, T_i; \theta_{w2r}) \\ &\quad + f_{w2r}(V_i, T_i^-; \theta_{w2r})] \end{aligned} \tag{7}$$

where n is the number of image-text pairs and parameter set θ_{w2r} is shared for both the positive and negative samples. This equation aims to ensure that the matching score of the positive pair (V_i, T_i) is greater than the negative pair (V_i, T_i^-).

For the positive pairs, it is supposed that the text T_i describes the semantics of the image V_i. The W2R attention model aims at aligning regions of image V_i with words of the text T_i. However, the calculating of the attention scores in Eq. (1) and Eq. (2) is inappropriate for learning the negative score Y_i^-, because there exists no alignments between the visual regions and text words of the negative pair (V_i, T_i^-). Hence, the way that attention scores are calculated (Eq. (1)) could misguide the model to obtain an incorrect matching. To address this problem, the following method is used to calculate the attention scores:

$$\alpha_{i,j}^k = \begin{cases} \dfrac{\exp(e_{i,j}^k)}{\sum_{j=1}^{D} \exp(e_{i,j}^k)} & input : (V_i, T_i) \\ \\ 1/D & input : (V_i, T_i^-) \end{cases} \tag{8}$$

In this way, the negative correspondences can be ignored and the positive alignments can be assigned with a reasonable value to $\alpha_{i,j}^k$. Note that we also utilize several CNNs to replace LSTM to build the W2R networks, but the experimental results show that W2R attention model with LSTM achieves slightly better performance than with CNNs.

B. Object to Words (O2W) Attention Network

On the other side, for a specific image object, there may exist some text words that are more semantically related to it. If the corresponding words are discovered for each object of the image, it is more effective to determine whether a text document is matched with the image. Therefore, O2W attention network is proposed to infer the semantic attention on the words attended by the image objects. Recently, it has been proved that semantic attentions are beneficial in many natural language processing related tasks, such as question answering and machine translation. In contrast to these works, our semantic attention model is built for cross-modal correlation learning.

To obtain the objects of an image V_i, the Faster-RCNN [43] is used to find the top P proposals. Then, the pre-trained deep CNNs is used to obtain the features $O_i = \{o_{i,1}, \ldots, o_{i,j}, \ldots, o_{i,P}\}^T \in \mathbb{R}^{P \times Q}$ of the P objects. For the text documents, the word embedding method is also used to obtain the representation as discussed above. Then, a bidirectional LSTM with the merging mode of "average" is used to generate the high level text features $S_i = \{s_i^1, \ldots, s_i^k, \ldots, s_i^L\}^T \in \mathbb{R}^{L \times B}$. For each object $o_{i,j}$, O2W attention network assigns a score $\beta_{i,j}^k \in [0, 1]$ to each word s_i^k based on the extent it relates to object $o_{i,j}$. Similar to W2R attention network, a $softmax$ function is used empirically to calculate $\beta_{i,j}^k$ as follows:

$$\beta_{i,j}^k = \frac{\exp(e_{i,j}^k)}{\sum_{k=1}^{L} \exp(e_{i,j}^k)} \tag{9}$$

where

$$e_{i,j}^k = \varphi((o_{i,j})^T W s_i^k + b) \tag{10}$$

is the unnormalized attention score measuring in what degree the word s_i^k is related to the object $o_{i,j}$. $\varphi(\cdot)$ is also the $tanh$ activation function. Similarly, $\beta_{i,j}^{:}$s normalize the attentions of all the words $\{s_i^k\}_{1 \leqslant k \leqslant L}$ drawn from object $o_{i,j}$. Then we use $\beta_{i,j}^{:}$s to adjust the attentive strength over different words. By weighted summing the the words for object $o_{i,j}$, the attended textual features can also be calculated as follows:

$$u_{i,j} = \sum_{k=1}^{L} \beta_{i,j}^k s_i^k, \quad U_i \in \mathbb{R}^{L \times M} \tag{11}$$

We denote the semantic attention process to automatically generate the attended text features as follows:

$$U_i = f_a(S_i, O_i; \theta_a), \quad U_i \in \mathbb{R}^{P \times B} \tag{12}$$

where θ_a is the weight parameters consisting of W and b in Eq.(2).

In contrast to the original text features which are shared by all objects, $u_{i,j}$ is more representative to show object-related words. We make concatenation of $u_{i,j}$ and $o_{i,j}$ and denote it as $c_{i,j} = [o_{i,j}; u_{i,j}]$, which can be treated as the output of the semantic attention process. Then each $c_{i,j}$ $(1 \leqslant j \leqslant P)$ is fed into multi-layer perceptron (MLP) to learn the matching score of each object and the corresponding text:

$$y'_{i,j} = f_m(c_{i,j}; \theta_m) \tag{13}$$

where f_m is the function to perform MLP and θ_m is the weight parameters.

In order to obtain the matching scores of the image with the corresponding text, we average the scores over the total objects in the image as follows:

$$Y'_i = \frac{1}{P} \sum_{1 \leq j \leq P} y'_{i,j}, \quad Y'_i \in \mathbb{R}^1 \tag{14}$$

By connecting Eq. (13) and Eq. (14), the process to obtain Y'_i from O_i and U_i is denoted as:

$$Y'_i = f_m(O_i, U_i; \theta_m), \quad Y'_i \in \mathbb{R}^1 \tag{15}$$

where f_m is the function combining MLP and the average calculation. To obtain an end-to-end process, we integrate the functions mentioned in Eq.(12) and Eq.(15) to calculate the matching score Y'_i from the multimodal input as:

$$Y'_i = f_{o2w}(V_i, T_i; \theta_{o2w}), \quad Y'_i \in \mathbb{R}^1 \tag{16}$$

where f_{o2w} is the integrating calculation of f_a and f_m, and θ_{o2w} is the set of parameters θ_a and θ_m. Similar to the W2R attention model, we employ siamese network structure to learn O2W attention network. To learn the matching scores Y'_i and Y'^{-}_i of the positive pair (V_i, T_i) and negative pair (V_i, T^-_i), the pairwise ranking loss can be formulated as follows:

$$\begin{aligned} L_2(\mathcal{V}, \mathcal{T}; \theta_{o2w}) &= \sum_{i=1}^{n} max[0, M - Y_i + Y'^{-}_i] \\ &= \sum_{i=1}^{n} max[0, M - f_{o2w}(V_i, T_i; \theta_{o2w}) \\ &\quad + f_{o2w}(V_i, T^-_i; \theta_{o2w})] \end{aligned} \tag{17}$$

where n is the number of image-text pairs and parameter set θ_{o2w} is shared among both the positive and negative calculations. The framework of the O2W attention network is shown in Figure. 4. Similar to the W2R attention network, we also re-calculate the attention weights to avoid misleading attention on the negative samples as Eq. (8).

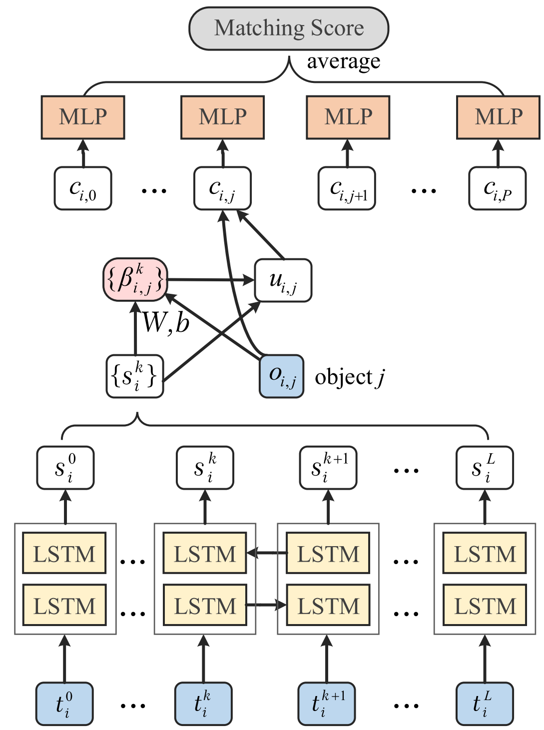

Fig. 4. The architecture of O2W attention network. The attention process produces $\{c_{i,1}, \ldots, c_{i,j}, \ldots, c_{i,P}\}$, where each $c_{i,j}$ is the concatenation of the attended textual feature $u_{i,j}$ and the corresponding object $o_{i,j}$. They are fed into MLP to obtain matching score for each object. The matching score of the image is the average over the objects.

C. A Bi-Directional Attention Model

As discussed above, two independent modules have been proposed to learn the matching score of image and text. The word to regions (W2R) attention network is learned to highlight the visual regions that are more related to the given words, while the object to words (O2W) attention network is learned to discover which words are more related to the given image object. The two attention networks can be combined to capture the bi-directional correlation between image and text more effectively in a complementary strategy. Intuitively, we propose a bi-directional spatial-semantic attention model to combine both of the W2R attention network and O2W attention network, by simultaneously optimizing them. Then the loss function is formulated as a summation of the hinge ranking losses of Eq. (7) and Eq. (17):

$$\begin{aligned} L(\mathcal{V}, \mathcal{T}; \theta_{w2r}, \theta_{o2w}) &= L_1(\mathcal{V}, \mathcal{T}; \theta_{w2r}) \\ &\quad + \lambda L_2(\mathcal{V}, \mathcal{T}; \theta_{o2w}) + \delta L_{reg}(\theta_{w2r}, \theta_{o2w}) \end{aligned} \tag{18}$$

where λ balances the importance of the W2R loss and O2W loss. L_{reg} is an L2-norm regularizer term to regularize the weight parameters in θ_{w2r} and θ_{o2w} to prevent overfitting. δ is a trade-off hyper-parameter.

In this fashion, the two directional attention networks are integrated into a holistic learning framework. The weighted sum of two ranking losses increases the overall complexity of the whole system. Both the W2R networks and the O2W networks usually require a careful selection of appropriate network structures. The hyperparameters λ and δ can be chosen through grid search. Since there are only two similar networks in the model which can be learned concurrently with GPUs in a unified objective function, it is not very time-consuming in practice. Note that we also attempt to utilize a fully connected layer to fuse the two attention networks, but the results show that this kind of integrating easily makes the model fall into local optimization.

The learning of the optimal image-text matching scores can be implemented by jointly minimizing the above loss function. Specifically, the stochastic gradient descent (SGD) with an adaptive learning rate is employed to train the model over the shuffled batches. After the training process, Eq. (6) and Eq. (16) can be combined to calculate the final matching score $MS(V_x, T_y)$ of a given pair of image (V_x) and text (T_y):

$$MS(V_x, T_y) = f_{w2r}(V_x, T_y; \theta_{w2r}) + \lambda f_{o2w}(V_x, T_y; \theta_{o2w}) \tag{19}$$

V. Experiments

In this section, we conduct extensive experiments on several public datasets to analyze the effectiveness of BSSAN in two tasks, i.e., image-to-text and text-to-image retrieval.

A. Dataset and Baselines

We test our model on the well-known Flickr30K dataset and MSCOCO dataset, which are both annotated with delicate descriptions. The details are described as follows:

- **Flickr30K** [45] contains 31,783 images collected from the Flickr website. We follow [1] which splits the dataset into training, validation, testing set with 29,783, 1,000, 1,000 images respectively. Each image is accompanied with 5 manually annotated sentences, resulting in $29,783 \times 5 = 148,915$ training pairs.
- **COCO** [46] contains 82,783 training images and 40,504 validation images. Each image is associated with 5 descriptions. It has $82,783 \times 5 = 413,915$ training pairs. Similar to [44], we use 82,783, 4,000, and 1,000 images to form training, validation, and test set respectively.

We compare our model with the state-of-the-art methods in the tasks of image to text retrieval and text to image retrieval. The baselines includes: the CCA-based methods, such as DCCA [18], CCA(FV HGLMM) [39], and CCA(FV GMM+HGLMM) [39]; and the ranking-based methods, such as mCNN [22], DSPE [40], RRF-Net(vgg) [47], DAN(vgg) [11], and 2WayNet [47]. Other methods, such as m-RNN [1] and DVSA [2], which are originally intended for image caption generating, are also compared in the two tasks.

To systematically validate the effectiveness, we implement four variant versions of BSSAN:

- **BSSAN**$_{w2r}$: it uses only the W2R attention network by removing O2W attention network from BSSAN.
- **BSSAN**$_{o2w}$: it uses only the O2W attention networks by removing W2R attention network from BSSAN.
- **BSSAN**$_{no\_reg}$: it is a simplification version of BSSAN by removing the L2-norm regularizer from BSSAN.
- **BSSAN**: it is a full implementation of our model that combines both of the attention networks and uses the L2-norm regularizer to prevent overfitting.

TABLE I
Neural Network Structure

Model	Type of layers	#Nodes in each layer
W2R networks	LSTM	1024
	MLP	1024-512-128-1
O2W networks	Bi-LSTM	512
	MLP	512-256-64-1

Note that the models proposed in DAN [11] and RRF-Net [47] use both ResNet-152 [48] and VGG-19 [49] networks to extract visual features. Since our model employs VGG-19 network to extract visual features, we only compare with DAN [11] and RRF-Net [47] using VGG-19 extractors for a fair comparison.

B. Experiment Preparation

For the visual input of the W2R attention network, we resize the images to 224×224 with channel RGB and employ the VGG-19 network [49] pre-trained on ImageNet dataset [50]. Similar to [3], the output of layer "conv5_4" is used as the region maps with the dimension of 196×512. Thus, each image has 196 candidate regions in total for W2R model. For the visual input of the O2W attention network, the top 10 object proposals extracted from each image using Faster-RCNN [43] are used. Each Object is represented with the 4096-dimension CNN feature vector extracted from the fc7 activation layer of the Faster-RCNN networks [43], which is fine-tuned with the combination of the PASCAL 2007 and 2012 train-val sets [51]. For the text features, the pre-trained word embedding built on GloVe [52] is employed to obtain a 300-dimensional vector. The zero padding is used to set a max length of the text descriptions with 30, and the sequences which are longer than 30 are truncated. The detailed structure of MLPs and LSTMs in the model are shown in Table I. We use $tanh$ activation function for MLPs and employ Batch Normalization [53] to reduce the internal covariate shift in the neural networks. The parameters λ and δ are experimentally set with 0.6 and 0.01.

During the process of training, SGD is employed for optimization of the model, which is set with the learning rate of 0.01 and the momentum of 0.9. All the experiments are performed on the workstation with 2×NVIDIA GeForce GTX 1080 and implemented with tensorflow[1] deep learning framework.

C. Experimental Results and Analysis

In this subsection, we show and analyze the experiment results of the two tasks on the two datasets. The trained

[1]https://github.com/tensorflow/tensorflow

TABLE II
BI-DIRECTIONAL IMAGE-TEXT RETRIEVAL RESULTS ON FLICKR30K

Methods	Image-to-Text				Text-to-Image			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
DCCA [18]	27.9	56.9	68.2	4.0	26.8	52.9	66.9	4.0
CCA (FV HGLMM) [38]	33.2	60.7	72.4	3.0	24.9	52.3	66.4	5.0
CCA (FV GMM+HGLMM) [38]	33.3	62.0	74.7	3.0	25.6	53.2	66.8	5.0
m-RNN-vgg [1]	35.4	63.8	73.7	3.0	22.8	50.7	63.1	5.0
DVSA (DepTree) [2]	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
DVSA (BRNN) [2]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	8.0
mCNN (ensemble) [22]	33.6	64.1	74.9	3.0	26.2	56.3	69.6	4.0
DSPE [39]	40.3	68.9	79.9	-	29.7	60.1	72.1	-
RRF-Net(vgg) [44]	42.1	-	-	-	31.2	-	-	-
DAN(vgg) [11]	41.4	73.5	82.5	2.0	31.8	61.7	72.5	3.0
2WayNet [44]	**49.8**	67.5	-	-	**36.0**	55.6	-	-
BSSAN$_{w2r}$	41.2	71.1	80.9	3.0	31.0	61.3	70.4	3.0
BSSAN$_{o2w}$	38.4	67.2	77.5	3.0	28.5	57.5	67.9	4.0
BSSAN$_{no_reg}$	43.1	73.9	82.8	2.0	32.3	61.7	72.3	3.0
BSSAN	44.6	**74.9**	**84.3**	**2.0**	33.2	**62.6**	**72.9**	**3.0**

TABLE III
BI-DIRECTIONAL IMAGE-TEXT RETRIEVAL RESULTS ON COCO

Methods	Image-to-Text				Text-to-Image			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
CCA (FV HGLMM) [38]	38.7	68.4	81.0	2.0	25.1	59.7	76.5	4.0
CCA (FV GMM+HGLMM) [38]	38.9	68.4	80.1	2.0	25.6	60.4	76.8	4.0
m-RNN-vgg [1]	41.0	73.0	83.5	2.0	29.0	42.2	77.0	3.0
DVSA [2]	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
DSPE [39]	50.1	79.7	89.2	-	39.6	75.2	86.9	-
2WayNet [44]	55.8	75.2	-	-	39.7	63.3	-	-
BSSAN$_{w2r}$	52.9	79.3	88.3	2.0	38.1	74.1	86.1	2.0
BSSAN$_{o2w}$	49.2	76.1	85.1	2.0	36.2	69.2	82.5	3.0
BSSAN$_{no_reg}$	54.7	81.9	89.7	1.0	40.7	75.5	87.9	2.0
BSSAN	**56.0**	**82.6**	**91.3**	**1.0**	**41.8**	**76.7**	**88.5**	**2.0**

models are used to predict the matching scores of a query sentence with the candidate images, and vice versa. Following [2] and [22], the metric R@K, i.e., the percent of queries whose top K matches contain at least one correct match, is used to evaluate the performance. The median rank (Med r) of the closest ground truth result is also used as the metric. The results for image and text matching on Flickr30K and COCO benchmarks are reported in Table II and III, respectively. Here, element denoted with "-" indicates that the results are missed in the corresponding papers.

From the results of Flickr30K in Table II, we can conclude that our BSSAN model surpasses the existing state-of-the-art methods and the variant versions in most cases. CCA-based methods such as DCCA and CCA(FV HGLMM) show relatively low performance due to that SGD cannot well solve the generalized eigenvalue problem. DSPE, RRF-Net(vgg), DAN (vgg), and 2WayNet are the most competitive methods for cross-modal retrieval. Among them, 2WayNet shows the powerful performance especially when considering the top result (R@1). However, our method obtains more obvious improvements on other metrics. The most similar method to our model is DAN (vgg). It builds two attention networks in two branches separately, which neglects the cross-modal correlation between image and text. Our model BSSAN outperforms DAN (vgg) in terms of all the metrics, which confirms the effectiveness of our model for image-text matching. From the results of BSSAN$_{w2r}$ and BSSAN$_{o2w}$, we can see that these two models with only one directional attention network almost achieve the performance of DSPE, which confirms the effectiveness of word to regions attention network and object to words (O2W) attention network. By comparing the results of BSSAN$_{w2r}$, BSSAN$_{o2w}$, BSSAN$_{no_reg}$ and BSSAN,

- College basketball player going up for a jump shot while another attempts to defend .
- A Miami Hurricanes basketball player has possession of the ball .
- One basketball player is shooting the ball while another one is trying to block his shot .
- The basketball player is number 21 from Miami .
- A college basketball game while a player is trying to pass the ball .

- A white and brown spotted dog , standing on its hind legs in front of a woman and a cake .
- Brown and white dog biting man 's leather jacket .
- A black-and-white dog bounds off the ground , all feet in the air , of a yellow field .
- A dog stands on his hind legs for a lady in purple .
- a white dog jumps into the air .

- A young boy in a red football jersey makes a large gesture as he stands in the midst of pumpkins .
- A boy is standing stretching his arms out whilst standing in a pumpkin patch .
- A boy in a pumpkin patch making a funny face with his arms in the air .
- A young boy with glasses and a blue sweatshirt holds a pumpkin that he picked from a pumpkin patch .
- A young boy holding a pumpkin in the middle of a pumpkin patch .

- group of people are wearing white T-Shirt are doing a collective work
- Five people are looking at something interesting through a glass .
- A group of teens and a group of adult men talk amongst themselves in an outdoor area .
- A group of people engaged in some sort of outdoor gathering or ritual .
- A group of young men are playing soccer .

Fig. 5. Four examples of the top 5 results returned by our approach from Flick30K in the task of Image-to-Text retrieval, where the correct descriptions are marked in green and the false descriptions are marked in red.

we can find that L2-norm and the fusion of two directional attention networks are effective to improve the performance of image-text matching.

On the other side, the experiment results on the dataset of COCO are reported in Table III. It also proves that BSSAN achieves remarkable improvement on various metrics. This is because that the attention networks can show the advantage for image-text matching when there is sufficient training data to capture the fine-grained cross-modal relations between the visual and textual contents. It is noteworthy that our model BSSAN outperforms all of the baselines on all metrics even compared with 2WayNet on R@1, which further validates the effectiveness of our model on the task of image-text matching.

To present a detailed analysis of the performance, we then select four test images randomly from the Flickr30K dataset to retrieve the textual descriptions. The 5 top-ranked text descriptions are shown in Figure 5. The right descriptions are marked in green, while the false ones are marked in red. Despite there exist false descriptions in the top-5 retrieved results, our model still captures the semantically related words in the false descriptions, e.g., “basketball player” of the first image and “white dog” of the second image. In the fourth image of the examples, the first text description “group of people are wearing white T-Shirt are doing a collective work” is mistakenly top ranked by our model. The possible reason is that both the W2R attention model and O2W attention model focus more on local information such as important words or discriminative regions. Therefore, words “people” and “white T-Shirt” in the first description have a strong alignment with the fourth image, which makes this description falsely ranked top by our approach. Meanwhile, Figure 6 presents two examples of the 5 top-ranked text-to-image search results for the models of BSSAN$_{w2r}$, BSSAN$_{o2w}$, BSSAN$_{no_reg}$ and BSSAN. As can be seen from the figure, BSSAN returns more reasonable images for the text queries. It further confirms the effectiveness of combining two directional attention networks for image-text matching.

D. Visualization of Attentions

To better understand the interpretability of our model, we take the advantage of the visual and semantic attention mechanism to visualize what the model “sees”. The visualization of word to regions (W2R) attention and object to words (O2W) attention is shown in Figure 7 and Figure 8 respectively.

For the W2R attention, the visualization approach is similar to [3]. That is, the image input to the convolutional network is resized to 224×224. Consequently, after 4 max-pooling layers, the dimension of the output on the top convolutional layer is 14×14. The attention scores are up-sampled with a multiplication factor of $2^4 = 16$ and then filtered with a Gaussian filter. Different from [3], we further draw a heat map for the up-sampled attention scores for brighter and more colorful visualization. Masked by the heat map with a transparency of 0.7, the original images are transformed to the final attended images. Therefore, the bigger the attention scores of the regions, the redder the regions are colored. From Figure 7, one can see that right attention is drawn from words to the corresponding image regions by the proposed attention model. For the description “A man stands in front of a pond with a green hill behind it”, the words “man”, “pond”, “green”, and “hill” are aligned to the right regions of the corresponding image. The word “stand” pays indistinct

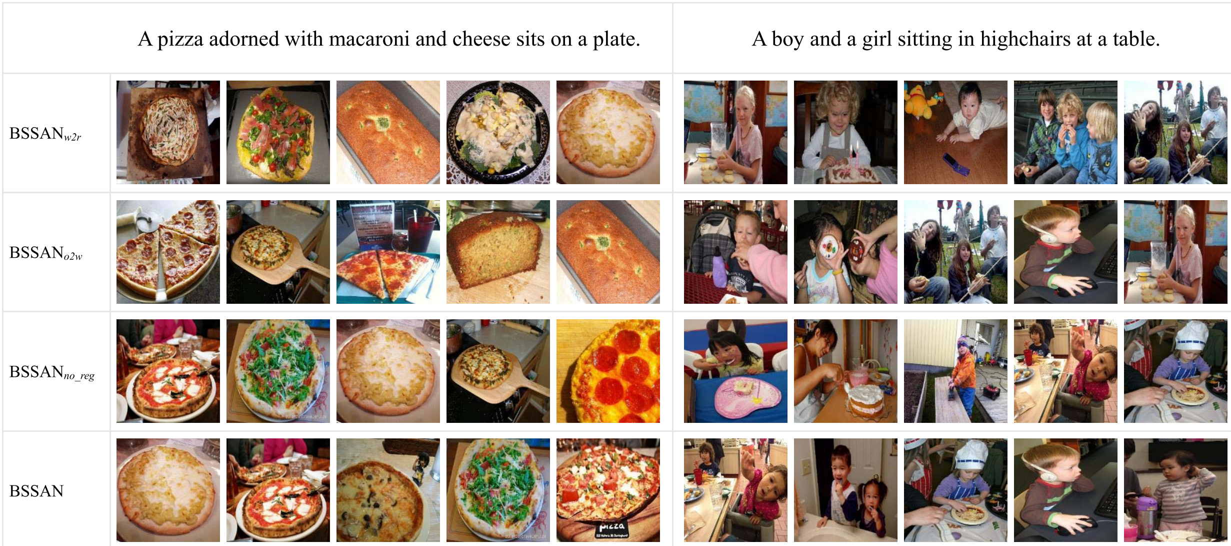

Fig. 6. Two examples of the text-to-image retrieval results with 5 top-ranked images on COCO, and the text queries are presented at the top.

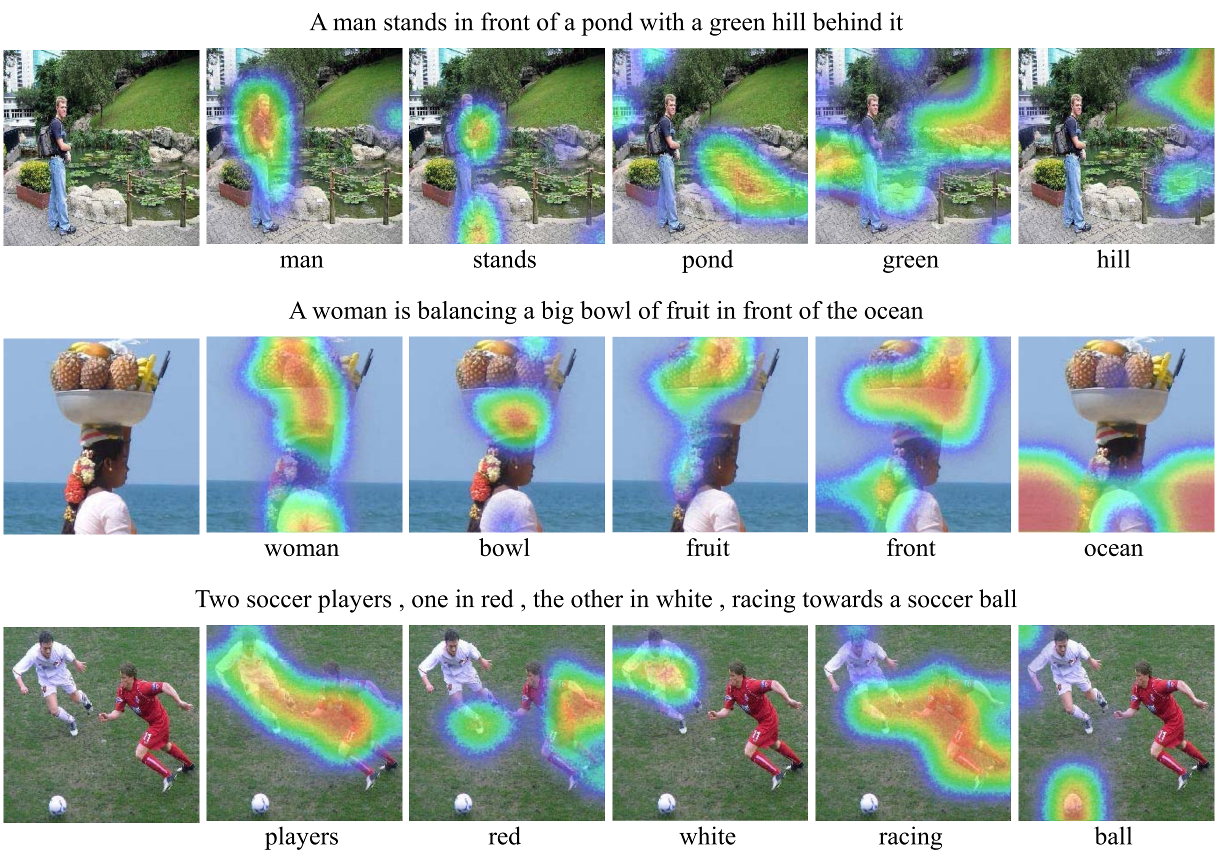

Fig. 7. Three examples of the learnt W2R attention.

attention on the image because it has no semantic information which can be manifested visually.

For the O2W attention network, it is easier to visualize because the attention weights are related to the textual words directly. Thus, we present a line chart for each image, where each line denotes the attention drawn from each object. For visualization simplicity, only the attention scores of 3 top-ranked objects are presented in the images in Figure 8. We can see that the objects in the image focus on the right semantic words by our attention model. For the first image and the sentence "a black and white dog carries a ball toy", the object in the yellow box pay strong attention on the words "black", "white", and "dog", and the attention from the object in blue box achieve a peak at the word "ball". However, the

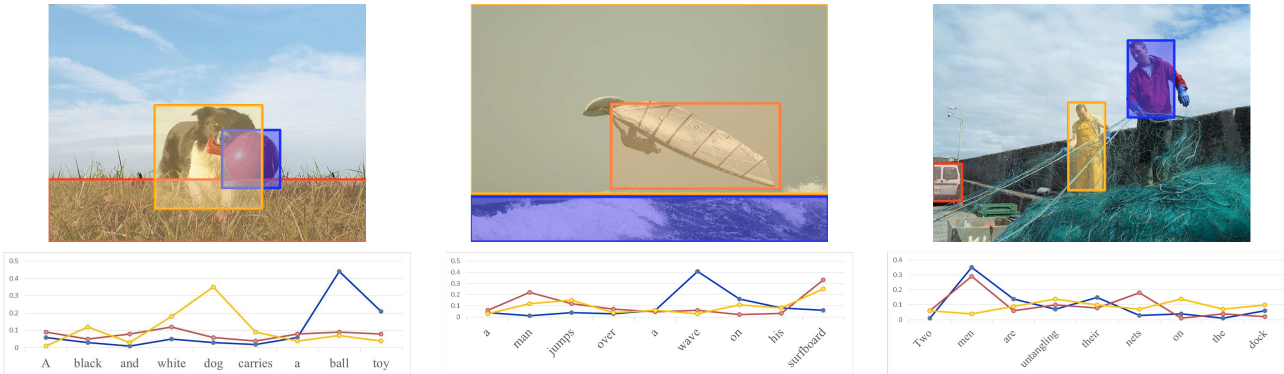

Fig. 8. Three examples of the learnt O2W attention.

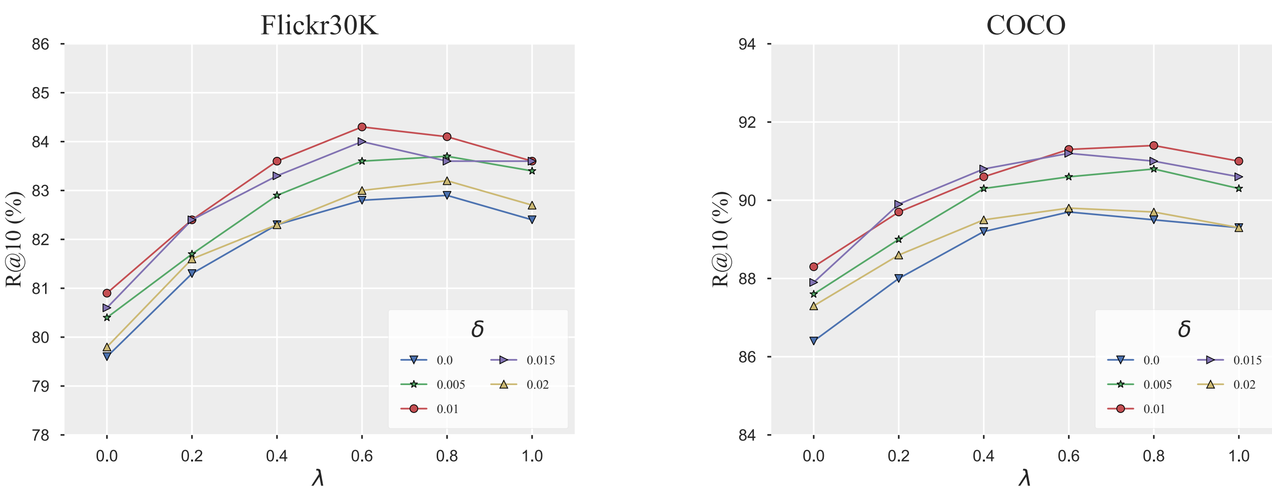

Fig. 9. parameter sensitivity study for λ and δ on Flickr30K and COCO.

object in the red box scatters the attention on every word equally, because it has no semantically-related words in the text description. This phenomenon will affect the performance of the O2W attention network, which results in that it performs worse than W2R attention network to some extent. Nevertheless, the matching score of O2W attention network is averaged across all the detected objects and provides complementary information to W2R attention network. Therefore, it still helps the joint model to learn a better matching score for the image-text pair.

The proposed attention mechanism makes the correlations between the image and the corresponding description interpretable and explicit. The W2R attention networks find the most related image regions for each word, while the O2W attention model discovers the semantically-related words to each object in the image. Combining the bi-directional attentions can provide a richer interpretation of the model and thus improve the performance.

E. Parameter Sensitivity Analysis

In order to evaluate in what extent does the parametrization affect the performance of BSSAN on image-text matching, we conduct the experiments on the two datasets by setting the parameters with different values. In the proposed model, we utilize λ to adjust the importance of two directional attention networks and δ to control the trade-off for the L2-norm regularizer.

We vary λ and δ to evaluate the performance of image-to-text retrieval on Flickr30K and COCO. The sensitivity curves are shown in Figure 9. One can see that setting the trade-off term is very necessary to prevent overfitting because the worst performance is achieved at $\delta = 0$. However, the performance also decreases when δ is assigned a large value. This may be due to the fact that too large value of the term affects the learning of image-text matching. Compared to δ, λ has a greater effect on the model performance. Specifically, the model depends entirely on W2R when $\lambda = 0$, which makes the semantic attention for image objects invalid totally. With the increase of λ, the model starts to put more and more emphasis on the optimization of O2W attention network. From Figure 9, it can be seen clearly that the model achieves optimal performance when $\lambda = 0.6$ and $\delta = 0.01$. The curves for text-to-image retrieval are similar to Figure 9, thus we do not present them.

VI. Conclusions and Future Works

In this paper, we explore to excavate the bi-directional and fine-granularity correlations between multimodal contents for image-text matching. In particular, a joint deep Bi-directional Spatial-Semantic attention Networks (BSSAN) is proposed. It uses two attention networks to model the bi-directional relations between image and text pairs. Then the two directional attention networks are unified into a holistic learning framework. We test BSSAN in the task of image-text matching on the benchmarks Flickr30K and COCO. The experiment results indicate that exploiting the correlation between image and text in the level of image objects and text words from two directions can improve the performance of image-text matching. Our approach is different from current image-text matching researches that mainly exploit one unidirectional relation or two types of relations separately.

Our approach can be easily generalized to other sequential data for cross-modal matching, e.g., voice and video. In the future work, we will investigate to integrate the two kinds of attention models together with a smoother framework rather than a weighted sum. we also want to fuse other information, e.g., the social relationship between images and the relationship between image owners, for more effective learning for image-text matching.

References

[1] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1412.6632

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3128–3137.

[3] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 2048–2057. [Online]. Available: http://jmlr.org/proceedings/papers/v37/xuc15.html

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4651–4659.

[5] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1612.01033

[6] A. Tariq and H. Foroosh, "A context-driven extractive framework for generating realistic image descriptions," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 619–632, Feb. 2017.

[7] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, vol. 9910, Amsterdam, The Netherlands: Springer, Oct. 2016, 2016, pp. 241–257.

[8] Y. Gao, M. Wang, Z. J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[9] Z. Liu, H. Li, W. Zhou, R. Zhao, and Q. Tian, "Contextual hashing for large-scale image search," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1606–1614, Apr. 2014.

[10] W. Zhou, H. Li, R. Hong, Y. Lu, and Q. Tian, "BSIFT: Toward data-independent codebook for large scale image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 967–979, Mar. 2015.

[11] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1611.00471

[12] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4540–4554, Oct. 2016.

[13] J. Shao, L. Wang, Z. Zhao, F. Su, and A. Cai, "Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval," *Neurocomputing*, vol. 214, pp. 618–628, Nov. 2016.

[14] W. Wang, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1610.03454

[15] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[16] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[17] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, vol. 28, Jun. 2013, pp. 1247–1255. [Online]. Available: http://jmlr.org/proceedings/papers/v28/andrew13.html

[18] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3441–3450.

[19] Z. Ma, Y. Lu, and D. P. Foster, "Finding linear structure in large datasets with scalable canonical correlation analysis," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, vol. 37, Jul. 2015, pp. 169–178. [Online]. Available: http://jmlr.org/proceedings/papers/v37/maa15.html

[20] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, Barcelona, Spain, Jul. 2011, pp. 2764–2770.

[21] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. 26th Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2013, pp. 2121–2129. [Online]. Available: http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model

[22] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2623–2631.

[23] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell.* San Francisco, CA, USA: AAAI Press, Feb. 2017, pp. 4133–4139. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14880

[24] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, vol. 9911. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 451–466.

[25] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. 29th Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 289–297. [Online]. Available: http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering

[26] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Austin, TX, USA, Nov. 2016, pp. 932–937. [Online]. Available: http://aclweb.org/anthology/D/D16/D16-1092.pdf

[27] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 21–29.

[28] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 842–850.

[29] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[30] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1702.05891

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[32] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Lisbon, Portugal, Sep. 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1166.pdf

[33] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. 28th Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 2953–2961. [Online]. Available: http://papers.nips.cc/paper/5640-exploring-models-and-data-for-image-question-answering

[34] H. Xue, Z. Zhao, and D. Cai, “Unifying the video and question attentions for open-ended video question answering,” *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5656–5666, Dec. 2017.

[35] F. Huang, X. Zhang, C. Li, Z. Li, Y. He, and Z. Zhao, “Multimodal network embedding via attention based multi-view variational autoencoder,” in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, K. Aizawa, M. S. Lew, and S. Satoh, Eds. Yokohama, Japan: ACM, Jun. 2018, 2018, pp. 108–116.

[36] L. Yang, Q. Ai, J. Guo, and W. B. Croft, “aNMM: Ranking short answer texts with attention-based neural matching model,” in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Indianapolis, IN, USA, Oct. 2016, pp. 287–296.

[37] T.-K. Kim, J. Kittler, and R. Cipolla, “Discriminative learning and recognition of image set classes using canonical correlations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[38] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, “Multi-view clustering via canonical correlation analysis,” in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, Montreal, QC, Canada, vol. 382, Jun. 2009, pp. 129–136.

[39] B. Klein, G. Lev, G. Sadeh, and L. Wolf, “Fisher vectors derived from hybrid Gaussian–Laplacian mixture models for image annotation,” *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1411.7399

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

[41] A. Karpathy, A. Joulin, and L. F. Fei-Fei, “Deep fragment embeddings for bidirectional image sentence mapping,” in *Proc. 27th Annu. Conf. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 1889–1897. [Online]. Available: http://papers.nips.cc/paper/5281-deep-fragment-embeddings-for-bidirectional-image-sentence-mapping

[42] Y. Peng, X. Huang, and J. Qi, “Cross-media shared representation by hierarchical learning with multiple deep networks,” in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, Jul. 2016, pp. 3846–3853. [Online]. Available: http://www.ijcai.org/Abstract/16/541

[43] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Montreal, QC, Canada, Dec. 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

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

[45] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014. [Online]. Available: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229

[46] T.-Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693. Zürich, Switzerland: Springer, Sep. 2014, 2014, pp. 740–755.

[47] A. Eisenschtat and L. Wolf, “Linking image and text with 2-way nets,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1855–1865.

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

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

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[51] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[52] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1162.pdf

[53] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, vol. 37. Jul. 2015, pp. 448–456. [Online]. Available: http://jmlr.org/proceedings/papers/v37/ioffe15.html

Feiran Huang received the B.Sc. degree from Central South University, Changsha, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include social media analysis and multi-modal data analysis.

Xiaoming Zhang received the B.Sc. and M.Sc. degrees in computer science and technology from the National University of Defense Technology, China, in 2003 and 2007, respectively, and the Ph.D. degree in computer science from Beihang University in 2012. Since 2012, he has become a Lecturer with the School of the Computer, Beihang University. He has published more than 30 papers, such as TMM, IEEE TRANSACTIONS ON CYBERNETICS, WWWJ, *Neurocomputing*, JIIS, *Signal Processing*, IJCAI2015, ICMR2015, SDM2014, WAIM2014, AAAI2013, and Coling2012. His major interests are social media analysis, image tagging, and text mining.

Zhonghua Zhao received the B.Sc. and M.Sc. degrees in computer science and technology from Shandong University, China, in 2005 and 2009, respectively, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences in 2013. He is currently with the National Computer Emergency Technical Team/Coordination Center of China. He has published more than 10 papers, such as WISE, WWWJ, IJDSN, and CCL 2016. His major interests are social media analysis and text mining.

Zhoujun Li received the M.Sc. and Ph.D. degrees in computer science from the National University of Defense Technology, China, in 1984 and 1999, respectively. Since 2001, he has become a Professor with the School of the Computer, Beihang University. He has published more than 150 papers on international journals, such as TKDE, *Information Science*, and *Information Processing and Management*, and international conferences such as SIGKDD, ACL, SIGIR, AAAI, IJCAI, SDM, CIKM, and WSDM. His research interests include the data mining, information retrieval, and database. He is a PC Member of many international conferences, such as SDM2015, CIKM2013, WAIM 2012, and PRICAI2012.