IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 2, 2020, accepted February 16, 2020, date of publication February 19, 2020, date of current version February 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2975036

Transfer Correlation Between Textual Content to Images for Sentiment Analysis

KE ZHANG [1], YUNWEN ZHU [1], WENJUN ZHANG [1,2], WEILIN ZHANG [3], AND YONGHUA ZHU [1]

[1] Shanghai Film Academy, Shanghai University, Shanghai 210000, China
[2] School of Communication and Information Engineering, Shanghai University, Shanghai 210000, China
[3] School of Computer Engineering and Science, Shanghai University, Shanghai 210000, China

Corresponding author: Yonghua Zhu (zyh@shu.edu.cn)

This work was supported in part by the National Key Research and Development Plan of China, under Grant 2017YFD0400101, and in part by the CERNET Innovation Project under Grant NGII20160325.

ABSTRACT In social media, images and texts are used to convey individuals' attitudes and feelings; thus, social media has become an indispensable part of people's lives. To understand social behavior and provide better recommendations, sentiment analysis on social media is helpful. One sentiment analysis task is polarity prediction. Although current research on visual or textual sentiment analysis has achieved quite good progress, multimodal and cross-modal analysis combining visual and textual correlation is still in the exploration stage. To capture a semantic connection between images and captions, this paper proposes a cross-modal approach that considers both images and captions in classifying image sentiment polarity. This method transfers the correlation between textual content to images. First, the image and its corresponding caption are sent into an inner-class mapping model, where they are transformed into vectors in Hilbert space to obtain their labels by calculating the inner-class maximum mean discrepancy (MMD). Then, a class-aware sentence representation (CASR) model assigns the distributed representation to the labels with a class-aware attention-based gated recurrent unit (GRU). Finally, an inner-class dependency LSTM (IDLSTM) classifies the sentiment polarity. Experiments carried out on the Getty Images dataset and Twitter 1269 dataset demonstrate the effectiveness of our approach. Moreover, extensive experimental results show that our model outperforms baseline solutions.

INDEX TERMS Correlation, cross-modal, transfer, sentiment analysis.

I. INTRODUCTION

As social media thrives, analyzing the sentiments in tweets has attracted increasing attention from researchers. On social media platforms such as Twitter and Facebook, people share their daily lives with images and short texts. To understand social behavior and provide better service to the users, a fundamental task is sentiment polarity classification.

Many tweets consist of two parts, an image and a text that is not long. Therefore, an accurate sentiment classifier must consider the two parts, and multimodal methods or cross-modal methods should be applied. One main challenge of multimodal or cross-modal sentiment analysis is that different modalities have individual semantic features. For tweets, the image and the text may not be correlated, which has a great impact on the classification accuracy.

Methods found to analyze sentiment have been explored in several studies. Existing sentiment analysis methods can be roughly divided into two categories [1] from a methodological perspective: traditional sentimental methods and hybrid sentimental methods.

Traditional sentiment analysis methods classify sentiment mainly by encoding the probability of words with sentiment relations, including keyword detection methods [2], classification and regression models [3], and semantic web methods [4]. Keyword detection approaches are the most widely used. Sentimental polarity is determined by counting the totality of sentimental words appearing in the corpus, such as happiness, sadness, and anxiety. However, such approaches fail to recognize ironic words. Such expositions are unsatisfactory because they only focus on counting sentimental

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

words in the corpus. Generally, people express their feelings such that more euphemistic and keyword detection methods are valid for a specific corpus only. Classification and regression models include support vector machines (SVMs) [5], [6], Bayesian reasoning [7], [8], and artificial neural networks (ANNs) [9]. By training a classifier with a corpus as input, the sentiment intensity of a keyword can be obtained. The above accounts fail to resolve the contradiction in the lack of semantic correlation, which means that other sentiment information related to the corpus is not considered. Methods based on the semantic web depend on ontology or knowledge graphs [10]–[12]. These methods no longer rely on keyword or word frequencies and use large-scale semantic knowledge graphs to mine hidden features between semantic concepts of corpora. Different from the other methods, semantic web models use semantic relations to reveal implicit sentiment and are often used on commercial websites. WordNet-Affect [13] and SenticNet [10] are representative large-scale sentiment knowledge graphs. These knowledge graph construction methods include iterative regression based on common sense graphs and inline regularization random walk algorithms. The error rate of these knowledge graphs is reduced by similarity comparison and the average maximum rate. The intensity value and polarity of all kinds of sentiment words are defined at the same time.

Hybrid sentiment analysis methods encode images and sentences into multidimensional distributed vectors, and then multimodal sentiment polarity is obtained by machine learning classification, which includes supervised, semisupervised and unsupervised learning [14]. The results of hybrid sentiment analysis are achieved mainly by image object feature extraction and multimodal fusion analysis. Among them, sentiment classification methods based on deep learning [15] and generative adversarial networks (GANs) [16] are the most popular.

However, visual sentiment semantics are high-level implicit semantics, which are different from explicit textual expressions; therefore, textual sentiment analysis methods realized by natural language processing have one-sidedness and uncertainty. The research to date has tended to focus on cross-modal image sentiment analysis.

Studies on image sentiment analysis have been mostly restricted to limited comparisons of image global features, and experimental data are rather controversial. In addition, different individuals have different attention and cognition to each region of an image. Accordingly, sentiment analysis on social media sites is still a difficult task. Then, scholars realized the mechanisms by which textual context underpins image semantics are not fully used, and cross-modal approaches were proposed. Cross-modal image sentiment analysis refers to methods that supply a gap between visual semantics and sentiment with textual context. Cross-modal studies offer some important insights into semantic fusion through transfer learning to map image features to textual labels. Then, the purpose of image annotation was achieved, and labels were used as prior knowledge in sentiment polarity classification. Nevertheless, the drawback of existing cross-modal image sentiment analysis methods is that they rely too much on mapped labels to understand the correlation between image content and textual context.

To increase the understanding of image sentiment by exploring the correlation between visual content and textual context, this paper proposes a novel cross-modal model for image sentiment analysis. First, a fine-tuning convolutional neural network (CNN) [17] and GloVe [18] are used to extract the features of an image and its caption. Second, the inner-class mapping model taking visual features as inputs calculates the inner-class maximum mean discrepancy (MMD) with corresponding textual features in the same Hilbert space to obtain their correlations, and then the correlations are represented as labels. Furthermore, the corresponding textual description is embedded into distributed representation by a class-aware attention-based gated recurrent unit (GRU) with redundant information filtered out. Third, an inline relationship between textual context and visual contents is obtained by an attention-based long short-term memory network (LSTM) to estimate the final image sentiment polarity.

The main contributions of this paper include the following:

(1) In this paper, a novel cross-modal image sentiment analysis model is proposed. This model extracts visual features and uses them as the attention weight parameter of LSTM to obtain the context image related in the corresponding textual description (caption). This model can be used to predict an image sentimental polarity by utilizing semantic correlation descriptions.

(2) Different from the existing cross-modal sentiment analysis methods, this paper proposes an inner-class mapping method based on unsupervised maximum mean discrepancy (MMD), which attempts to learn cross-modal mapping correlations between images and descriptions.

(3) The end-to-end sentiment analysis algorithm is implemented in this paper. The experimental results show that the precision, F1 and accuracy are improved, and the proposed model outperforms other state-of-the-art image sentiment analysis methods on the Twitter1269 dataset. The feasibility and effectiveness of the model are also validated by a case study.

The remainder of this paper proceeds as follows. In Section 2, the layout methods of previous sentiment analysis research are reviewed, along with how representative cross-modal methods work. Section 3 describes the methodology used for this study. Section 4 analyses the experimental results and discusses some data and examples. Finally, Section 5 gives a brief conclusion.

II. RELATED WORK

Sentiment analysis is becoming increasingly important due to the rise of actual needs in social media platforms. There are some volumes of published studies describing the role of cross-modal research. Much of the cross-modal research has simply focused on identifying and evaluating algorithms based on text features or visual features.

A. TEXTUAL SENTIMENT ANALYSIS

Previous textual feature-based studies used in cross-modal approaches have explored the connection between textual features and sentiment, such as topic word detection models [19] and sentence grammar layer models [20]. These methods have achieved remarkable results and provided a significant reference for image sentiment analysis. Approaches in the literature can be classified into two categories: aspect-based sentiment analysis (ABSA) and targeted sentiment analysis (TSA).

1) ASPECT-BASED SENTIMENT ANALYSIS

The task of ABSA is to classify sentiment polarity by analyzing words to obtain an aspect. For example, ‘‘this computer is very expensive’’, and ‘‘price’’ is the aspect. The main challenge of ABSA is to classify the sentiment polarity of a compound sentence with multiple aspect words.

Zainuddin *et al.* [21] proposed an aspect-based hybrid feature selection method for Twitter. Wang *et al.* [22] analyzed sentiment by merging the attention in a multilayer neural network. For each word in a sentence, the attention weight shows the most pivotal word in a sentence, and the association degree of a given aspect word is obtained after the dot product. The experimental results show that this method can reduce the training loss caused by the recurrent neural network (RNN), and the accuracy of this multilayer neural network classifier is much higher than that of a single output layer classifier. Liu *et al.* [23] proposed a model combining regional CNN and LSTM; this model retains content information and time-series relationships between sentences in the whole comment without additional dependency analysis.

2) TARGETED SENTIMENT ANALYSIS

TSA methods extract specific target words in a sentence to analyze the relationship between the target word and some sentiment words through LSTM, such as target dependency LSTM (TDLSTM) and target correlation LSTM (TCLSTM) [24]. TDLSTM matches the hidden output layer of the bi-LSTM encoder with the target word to obtain sentence polarity. TCLSTM is extended by TDLSTM, which encodes each input word with the target word to obtain sentence distributed representation.

Attention is also suitable for TSA. Tang *et al.* [25] used RNN with multilayer attention weights to obtain classification results under supervised learning. This method improves the weights of important words by multihop training. Because only the weight of one word is increased, this method is unsuitable for sentences with multiple important words. Lu and Wu [26] constructed sentiment dictionaries to automatically extract sentiment words from a corpus to sentiment classifiers. Then, SVM was used to determine the final polarity. Bin *et al.* [27] proposed a new target-specific sentiment analysis approach based on a multiattention CNN. This method can take parallel text as input and greatly decrease the loss during training time. Additionally, this method can effectively compensate for the deficiency of a single attention layer.

B. VISUAL SENTIMENT ANALYSIS

Visual sentiment analysis is carried out by designing a polarity classifier to analyze visual features [28], [29]. Previous studies have established models including low-level feature extraction [30], [31], the semantic feature model [32], [33], and the deep learning framework [34], [35]. These approaches are mainly focused on low-dimensional feature extraction, such as color histograms, and the most typical method is human facial emotion recognition [36]. Human facial emotion is the most obvious sentiment symbol and is easy to identify. However, this kind of method is not applicable in other domains because of the semantic gap between low-level and high-level features.

With the improvement of deep learning, You *et al.* [37] used a pretrained domain transfer learning approach to analyze sentiment. Ahsan *et al.* [38] extracted an intermediate visual representation of social event images based on the visual attributes that occur in images going beyond sentiment-specific attributes. Song *et al.* [39] proposed a multilayer attention network to capture the saliency of the image content region, and sentiment polarity was classified according to the content with saliency. However, this method is effective only for simple images, especially for images containing only one object as content. Dong *et al.* [40] proposed four shared networks that receive multiple instances as inputs and are connected by a novel loss function consisting of pair-loss and triplet-loss to examine the potential connections among training instances. This method achieves excellent performance on object tracking. In fact, the sentiment analysis application scenario is always complicated.

C. MULTIMODAL SENTIMENT ANALYSIS

Due to the lack of direct mappings from visual semantics to sentiment, social media networks provide other abundant types of information, such as image captions and videos. Several studies [41]–[43] have used multimodal data to construct sentiment classifiers.

It is difficult to directly map images to sentiment. Wollmer *et al.* [44] and Kumar *et al.* [45] proposed a model to recognize facial emotion by multimodal feature fusion. Poria *et al.* [46], [47] proposed sharing state parameters from a CNN model with multikernel learning (MKL). Zadeh *et al.* [48] proposed a tensor fusion method. Byrne *et al.* [49] proposed employing simultaneous derivation to facial emotion recognition, but the essence of this method is still based on statistics. Xu *et al.* [50] proposed a hierarchical deep fusion model to explore the cross-modal correlations among images, texts, and social links, which can learn comprehensive and complementary features for more effective sentiment analysis. Their work is interesting and novel. The drawback of this model is that it is only for specific links, and overall, these links are unreliable on social media. Borth *et al.* [51], Maurya *et al.* [52], and Li *et al.* [53] proposed using sentiment-related adjective-noun pairs (ANPs). By means of ANP extraction, a visual sentiment ontology (VSO) was constructed, such as Sentibank [32] and SenticNet [10]. Similarly, Teng-Jiao *et al.* [54] proposed an object-sentiment pair extraction method based on middle-level semantic and grammatical analysis. Most studies on ANPs and VSO have high accuracy.

D. CROSS-MODAL SENTIMENT ANALYSIS

Different from the multimodal approach, cross-modal image sentiment analysis attempts to construct a mapping model based on transfer learning. To solve the lack of labeled training data, transfer learning spreads the knowledge from the source domain to the target domain by finding the similarity rules between data in two domains. Tsai *et al.* [55] proposed heterogeneous transfer from one modality to another, which is still a one-to-one transfer paradigm. Huang *et al.* [56] performed knowledge transfer between two domains, and models in two domains both share the same parameters. Schmitter *et al.* [57] proposed mapping semantic labels by estimating the probability distributions of the source domain to the target domain. Ji *et al.* [58] proposed a novel bilayer multimodal hypergraph learning (Bi-MHG) for robust sentiment prediction of multimodal tweets. Van Opbroek *et al.* [59] used kernel learning to calculate the weight parameters of image segments for polarity classification. Wu *et al.* [60] proposed a multilabel image annotation approach through sharing training structures. Huang *et al.* [56] calculated the mapping distance between the image feature and the semantic label through the MMD distance square.

Although cross-modal sentiment analysis methods solve the problem of insufficient data through transfer learning, these sentiment classifiers have ignored the semantic inner relation in textual descriptions. Attention [61], [62] weights are available parameters for analyzing inner relations, but attention is only widely used in image object detection.

Inspired by the work [56] on learning joint visual and textual models, this paper relies on the MMD distance to embed similarity between images and description for image annotation. This is different from previously mentioned works, all of which were proposed to extract the descriptive context of a related image for sentiment analysis by a class-aware IDLSTM.

III. METHOD

In this section, we propose our approach for cross-modal sentiment analysis and present its detailed explanation. The architecture of our approach is shown in Figure 1. To finish the task, an image and its corresponding description are fed into our model as the input. First, in stage (a), the image and its caption are processed separately. The image goes through a fine-tuned CNN, and the visual features of the sample are then fetched. The caption text is transformed into a sequence of vectors through word embedding methods, which are the textual features of the sample. In stage (b), a joint mapping model is performed to map the visual and textual features into the same Hilbert subspace. This is the key procedure of our approach on which the eventual classification accuracy greatly depends. Finally, in stage (c), class-aware sentence representation (CASR) is carried out, and the inner-class dependency LSTM (IDLSTM) is responsible for the sentiment polarity classification. Section III. A will provide more information about how the visual and textual features are extracted in stage (a). Section III. B will explain in detail the joint mapping model of stage (b). Section III. C will illustrate the procedure of CASR and IDLSTM.

A. VISUAL AND TEXTUAL FEATURE EXTRACTION

The input image and its caption are processed separately to obtain visual features and textual features. The image is analyzed by a fine-tuned VGGNet-16 [17], and the descriptive sentences are transformed into vectors through GloVe [18].

In the fine-tuned VGGNet-16, each convolutional feature map f_i corresponds to one specific region in the image, where N is the number of feature maps, and D_I is the representation dimension for each region. Specifically, the extracted image feature maps F_I from a raw image I through VGGNet-16 are denoted as follows:

$$F_I = CNN_{VGGNet}\,(I) \in \mathbb{R}^{D_I \times N} \tag{1}$$

$$F_I = [f_1, \ldots, f_i, \ldots, f_N]\,, f_i \in \mathbb{R}^{D_I} \tag{2}$$

In this case, the input image is fed with a resolution of 225×225 into VGGNet-16. The output of the Conv5_3 layer is 15×15, and the dimension N is 512.

The description of a given image is represented as $S = [w_1, \ldots, w_i, \ldots, w_L]$, where w_i is each word of a sentence, and L is the maximum number of words in the description. Each word w_i is embedded as a 300-dimensional GloVe [18] word vector $v_i \in \mathbb{R}^{300}$, and the sentence is represented as $V_s = \{v_1, \ldots, v_i, \ldots v_L \in R^{300\times L}$.

B. JOINT MAPPING MODEL

Transfer learning is applied in our approach to construct the correlation between image objects and labels. Previous cross-modal approaches learn global domain shifts by projecting all visual and textual features in both domains into a single subspace, which causes the absence of intra-affinity within classes [63]. To address the problem, our approach utilizes the intra-affinity of classes shared by both visual and textual domains, and an inner-class mapping model (IMM) is therefore proposed.

In the training stage, there are two domains, a labeled source domain $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$ and an unlabeled target domain $\mathcal{D}_t = \{(x_j)\}_{j=1}^{n_t}$, where $x_i, x_j \in \mathbb{R}^{d_1}$ are visual features. The source and the target visual vector, whose vector spaces are denoted by $\mathcal{X}_s, \mathcal{X}_t$, respectively, should share the same vector space $\mathcal{X}$ but are subject to different distributions. Similarly, the source and the target textual vector, whose vector spaces are denoted by $\mathcal{Y}_s, \mathcal{Y}_t$, share the same vector space $\mathcal{Y}$ but are subject to different distributions where $y_i \in \mathbb{R}^{d_2}$ represents the textual features. During the training process, the domain shifts with an increase in the number of image samples. Therefore, this paper assumes that when the marginal distribution $P(X_s) \neq P(X_t)$ and the conditional distributions $Q(Y_s|X_s) \neq Q(Y_t|X_t)$, a unified transform function $\phi(\cdot)$ exists.

The similarity of the visual and textual features in both $\mathcal{D}_s$ and $\mathcal{D}_t$ is the primary consideration for transfer learning. In our approach, maximum mean discrepancy (MMD) [64] is utilized to learn the potential features in the reproducing kernel Hilbert space (RKHS). The MMD between domains is formulated as:

$$D\left(\mathcal{D}_s, \mathcal{D}_t\right) = \left\| \frac{1}{n_s} \sum\nolimits_{x_i \in \mathcal{D}_s} \phi\left(x_i\right) - \frac{1}{n_t} \sum\nolimits_{x_i \in \mathcal{D}_t} \phi\left(x_j\right) \right\|_{\mathcal{H}}^2 \tag{3}$$

where $\mathcal{H}$ is the RKHS, and $\phi(\cdot)$ is the unified transform function. Then, the original sample feature vectors are mapped to RKHS.

To make use of the intra-affinity of classes shared by visual and textual features, this paper improves transfer component analysis (TCA) [65] and proposes an interclass distance. The distance between classes is measured as:

$$\begin{aligned} &D\left(\mathcal{D}_s, \mathcal{D}_t\right) \\ &= \frac{1}{n_s n_t} \sum\nolimits_{c=1}^{C} \left\| \frac{1}{n_s^c} \sum\nolimits_{x_i \in \mathcal{D}_s^c} \phi\left(x_i\right) - \frac{1}{n_t^c} \sum\nolimits_{x_j \in \mathcal{D}_t^c} \phi\left(x_j\right) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{4}$$

where $c \in \{1, 2, \ldots, C\}$ denotes classes, $\mathcal{D}_s^c, \mathcal{D}_t^c$ represent feature sets belonging to class c in the source and target domains, and n_s^c, n_t^c are the number of feature vectors belonging to class c in the source domain and target domain, respectively. The factors in (4) are used to average the MMD distances of all features in the same class and prevent them from being influenced by individuals.

The TCA approach converts the data of the two domain spaces to a new Hilbert space to reduce the difference and solves the semidefinite programming problem (SDP) by constructing the kernel matrix. Then, the MMD distance and the kernel matrix are written as:

$$D\left(\mathcal{D}_s, \mathcal{D}_t\right) = tr(KL_c) \tag{5}$$

$$K = \langle \phi(x_i), \phi(x_j) \rangle = \phi\left(x_i\right)^T \phi(x_j) \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)} \tag{6}$$

where L_c is an MMD matrix, and K is a kernel matrix constructed by the inner product of the mapping. A transformation matrix $W \in \mathbb{R}^{(n_1+n_2)\times m}$ converts the data from the original space into the RKHS, where $m \ll d$ is the dimension of RKHS. Equation (4) is converted by TCA to:

$$D\left(\mathcal{D}_s, \mathcal{D}_t\right) = \frac{1}{n_s n_t} \sum\nolimits_{c=1}^{C} tr(W^T KL_c KW) \tag{7}$$

For the trace optimization problem, the minimum MMD distance needs to be determined by kernel tricks, and the solution objective of (7) is rewritten as:

TABLE 1. Inner-class mapping model.

```
ALGORITHM 1.
Input: Source Domain $\mathcal{D}_s = \{X_s, Y_s\}$; Target Domain $\mathcal{D}_t = \{X_t\}$;
Output: Image feature vectors of Target Domain $\{Y_t\}$;
$K \leftarrow (6)$
$L_c \leftarrow (9)$
Repeat:
    Update $W$ by (11)
    Update $X_s$ by $W$
    Update $K$ by (6)
    Update $L_c$ by (9)
Until Converge
return $\{y_t\}$
```

$$\begin{aligned} &\min_W \sum\nolimits_{c=1}^{C} tr\left(W^T KL_c KW\right) + \Theta tr\left(W^T W\right) \\ &s.t. W^T KHKW = I \end{aligned} \tag{8}$$

$\sum_{c=1}^{C} tr\left(W^T KL_c KW\right)$ calculates the MMD distance of the feature vectors in each class in (8). $\Theta tr\left(W^T W\right)$ represents a regularization term, and Θ is the trade-off factor to ensure that the model is well defined. Constraints $W^T KHKW = I$ are used to maintain the data variance, where $I \in \mathbb{R}^{m\times m}$ is an identical matrix:

$$(L_c)_{ij} = \begin{cases} \frac{1}{\left(n_1^c\right)^2} x_i, & x_j \in D_s^c \\ \frac{1}{\left(n_2^c\right)^2} x_i, & x_j \in D_t^c \\ -\frac{1}{n_1^c n_2^c} & \begin{cases} x_i \in D_s^c, & x_j \in D_t^c \\ x_i \in D_t^c, & x_j \in D_s^c \end{cases} \\ 0 & otherwise \end{cases} \tag{9}$$

The Lagrange multiplier Φ is used to solve (8) as:

$$\begin{aligned} L = tr&\left(W^T K \sum_{c=1}^{C} L_c K^T W\right) \\ &+ \Theta tr\left(W^T W\right) \\ &+ tr(\Phi(I - W^T KHKW)) \end{aligned} \tag{10}$$

Equation (9) is nonconvex and can be finally formalized as a generalized eigendecomposition problem by setting derivative $\frac{\partial L}{\partial W} = 0$:

$$\left(K \sum\nolimits_{c=1}^{C} L_c K^T + \Theta I\right) W = KHK^T W \Phi \tag{11}$$

Finally, the minimum m of the generalized eigendecomposition in (11) is taken to obtain the transformation matrix W. The optimized algorithm is shown in Table 1, and the locally optimal solution is obtained when m converges by iterations.

C. SENTIMENT CLASSIFICATION

In this section, a sentiment classifier is introduced. The sentiment classifier is composed of two models: a class-aware

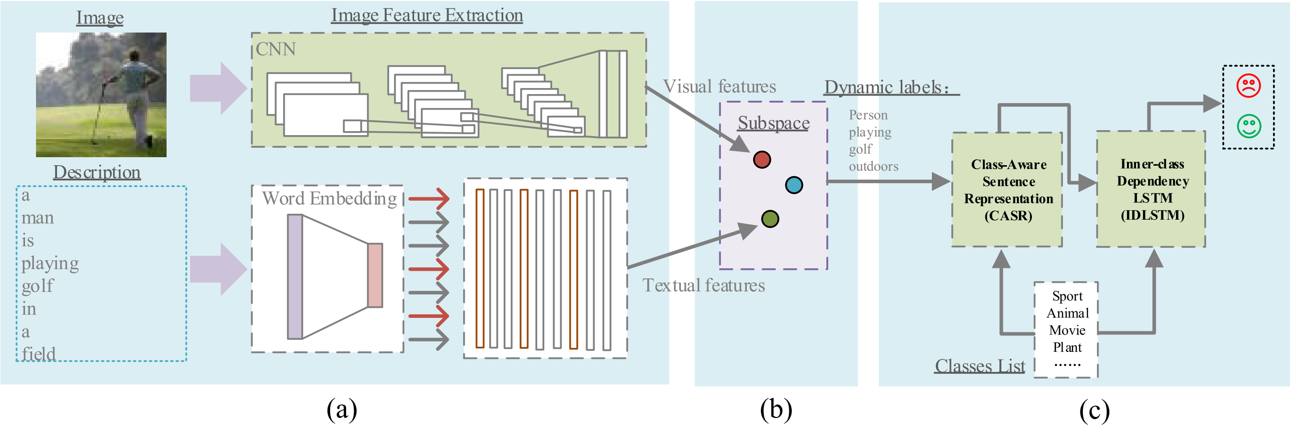


FIGURE 1. The architecture of the proposed model.

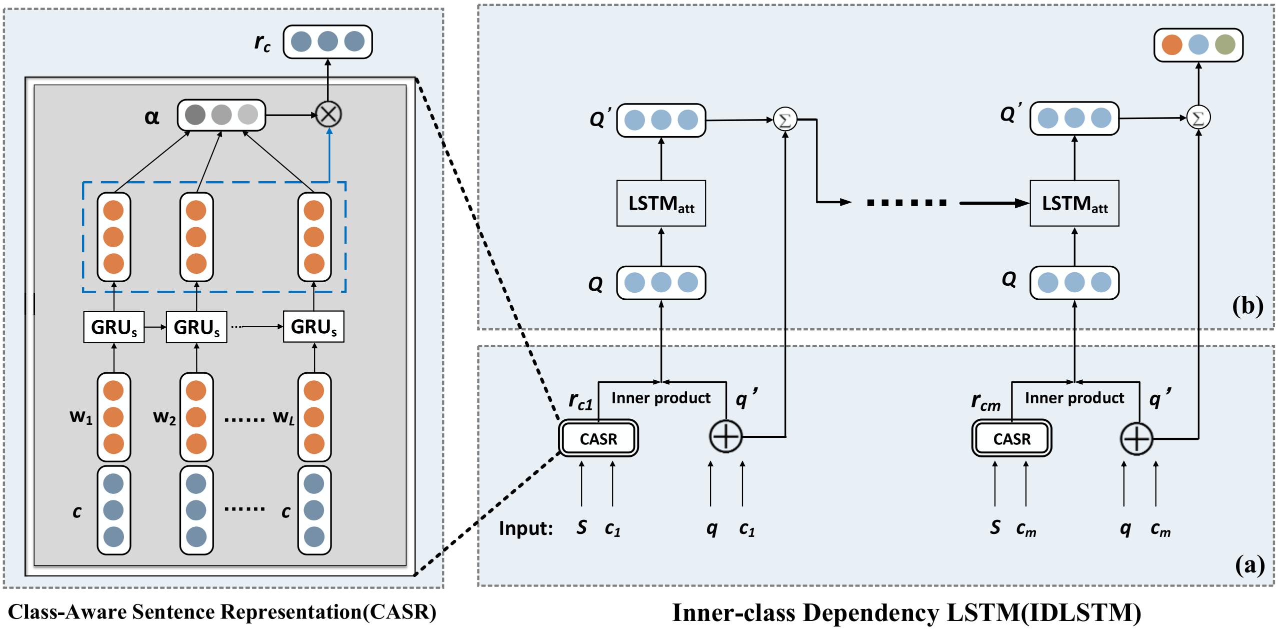


FIGURE 2. The architecture of image sentiment classification.

sentence representation (CASR) and an inner-class dependency LSTM (IDLSTM). CASR concatenates all words in the descriptive sentences with their given class representation, and IDLSTM is an attention-based LSTM designed for modeling the dependence of the target class with the other classes in descriptive sentences. The architecture of the proposed image sentiment classifier is shown in Figure 2.

1) CLASS-AWARE SENTENCE REPRESENTATION

The description of a given image is represented as $S = [w_1, \ldots, w_i, \ldots, w_L]$, where w_i represents each word of a sentence in the description, and L is the maximum number of words in the description. Every word in the sentence belongs to a class and class set $C = [c_1, \ldots, c_i, \ldots, c_M]$, where M is the maximum number of classes. Considering the deviation of different people to the same thing, such as "puppy" and "dog", classes are utilized to reduce the loss of deviation. Each word w_i and class c_i is embedded as a 300-dimensional (D) GloVe [18] word vector. Then, the description of the given image is represented as $S \in \mathbb{R}^{L\times D}$, and the class word is represented as $c_i \in \mathbb{R}^{D}$. Using the literature [31] for reference, each word w_i is associated with a given class c_i to form a sequence, and the sentence S is represented as:

$$S_{c_i} = \{w_1 \oplus c_i, w_2 \oplus c_i, \ldots, w_L \oplus c_i \in \mathbb{R}^{L\times 2D} \quad (12)$$

The distributed representation S_{c_i} is then fed to a GRU for context propagation, followed by an attention layer to obtain the class-aware sentence representation. The GRU is formulated as follows:

$$z = \sigma\left(x_t U^z + s_{t-1} W^z\right) \tag{13}$$

$$r = \sigma\left(x_t U^r + s_{t-1} W^r\right) \tag{14}$$

$$h_t = tanh\left(x_t U^h + (s_{t-1} * r) W^h\right) \tag{15}$$

$$s_t = (1 - z) * h_t + z * s_{t-1} \tag{16}$$

where h_t is the output of the hidden layer, s_t is the cell state at time t, σ and $tanh$ are activation functions, z is the update gate, and r is the reset gate. This step is represented as follows: The whole step of GRU is represented as $R_{c_i} = GRU_s(S_{c_i})$, where $R_{c_i} \in \mathbb{R}^{L \times D_s}$, $U_s^z \in \mathbb{R}^{2D \times D_S}$, $W_s^z \in \mathbb{R}^{D_S \times D_S}$, $U_s^r \in \mathbb{R}^{2D \times D_S}$, $W_s^z \in \mathbb{R}^{D_S \times D_S}$, $U_s^h \in \mathbb{R}^{2D \times D_S}$, $W_s^h \in \mathbb{R}^{D_S \times D_S}$.

The sentiment polarity and intensity of each word are different. To highlight the sentimentally relevant words to class c_i, an attention layer is added to capture the weight of each word:

$$z = R_{c_i} W n_s + b_s \tag{17}$$

$$\alpha = softmax\,(z) \tag{18}$$

$$r_{c_i} = \alpha^T Rnn_{c_i} \tag{19}$$

where z is the output of the update gate, $z = [z_1, z_2, \ldots, zn_L] \in \mathbb{R}^{L \times 1}$, $softmax\,(x) = \left[\frac{e^{x_1}}{\sum_j e^{x_j}}, \frac{e^{x_2}}{\sum_j e^{x_j}}, \ldots, \frac{e^{x_j}}{\sum_j e^{x_j}}\right]$, attention weight $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_L] \in \mathbb{R}^{D_s \times 1}$, The distributed representation combined attention weight is formulated as $r_{c_i} \in Rnn^{D_s}$, $W_s \in \mathbb{R}^{D_s \times 1}$, and b_s is a scalar.

2) INNER-CLASS DEPENDENCY LSTM

The labels of a given image are obtained, as is the class-aware sentence representation r_{c_i} in former sections. The aim of this section is to reinforce the descriptive context related to an image. This paper proposes an LSTM to model the dependency of label vectors with other word vectors in descriptive sentences by increasing the weight of words associated with image labels. The sentiment classifier IDLSTM consists of 2 partitions, as shown in Figure 2.

The mapped image label was normalized as the query q for further memory networking. To reduce the loss caused by the inconsistency between the label and the description in the morphology, class c_i is concatenated to image label q as $q' = q \oplus c_i \in \mathbb{R}^{2D}$. The distributed representation Q is supplied as the memory slot in Figure 2(a).

$$Q = q' r_{a_i}^T \tag{20}$$

$$\beta = softmax(Q) \tag{21}$$

where $Q = [Q_1, Q_2, \ldots, Q_M] \in \mathbb{R}^{M \times 1}$ and attention weight $\beta = [\beta_1, \beta_2, \ldots, \beta_M] \in \mathbb{R}^{M \times 1}$. Each β_i is a strength value of the match between each word and a label.

TABLE 2. An attention-based IDLSTM.

```
Algorithm 2.
        Input: Training set $(I, D)$; $I$:images; $D$: description;
        Output: The image sentiment prediction;
        /* Initialization */
        for $i$: $[1, M]$ do
        $r_{c_i} \leftarrow \boldsymbol{CDSR}(S, c_i)$
        /*Image annotation*/
        $f \leftarrow (1)$;
        $q \leftarrow$ CrossModelFeatureMap$(f)$;    // Algorithm 1
        $q' \leftarrow q \oplus c_i$
        /*LSTM training stage*/
        $Q \leftarrow (20)$
        $Q' \leftarrow (22)$
        for $i$: $[1, H]$ do            // $H$: $Hops$
                $Q \leftarrow (20)$
                $\beta \leftarrow (21)$
                $Q' \leftarrow (22)$
                $o \leftarrow (23)$
            $q' \leftarrow (24)$
        /*Classification*/
        $\hat{y} \leftarrow (25)(26)$
        return $\hat{y}$
        Procedure $\boldsymbol{CDSR(S, c)}$
        $S_c = (12)$
        $R_c = GRU_s(S_c)$
        $z = (17)$
        $\alpha = (18)$
        $r_c = (19)$
        return $\boldsymbol{r_c}$
```

Each word in the sentence is represented by the corresponding class-aware sentence representation. Considering that the memory value is usually too small and easy to forget, this paper uses attention-based LSTM $LSTM_{att}$ with size D_o to predict the correct classification of these words. As shown in Figure 2(b)

$$Q' = LSTM_{att}\,(Q) \tag{22}$$

where the parameters of $LSTM_{at}$ are $U_a^z \in \mathbb{R}^{D_s \times D_o}$, $W_a^z \in \mathbb{R}^{D_o \times D_o}$, $U_a^r \in \mathbb{R}^{D_s \times D_o}$, $W_a^r \in \mathbb{R}^{D_o \times D_o}$, $U_a^h \in \mathbb{R}^{D_s \times D_o}$, and $W_a^h \in \mathbb{R}^{D_o \times D_o}$. The response vector o is obtained by summing output vectors in Q', weighted by the relatedness measures in β:

$$o = \beta^T Q n' \tag{23}$$

where $o \in R^{D_o}$.

In the final stage, distributed representation q related to the image is added to the memory output of o to generate the predicted value.

$$q'_{(h+1)} = q'_{(h)} + o \tag{24}$$

$$\mathrm{P} = \mathrm{softmax}[\left(q'_{(h+1)}\right) W_{smax} + b_{smax}] \tag{25}$$

$$\hat{y} = \underset{i}{\mathrm{argmax}}\,(P[i]) \tag{26}$$

where $W_{Smax} \in \mathbb{R}^{D_o \times C}$, $b_{smax} \in R^C$, and the maximal value of $\hat{y}$ is taken as the prediction. Table 2 shows the algorithm of all steps.

3) LOSS FUNCTION

In this paper, the memory network is trained for 30 epochs using cross entropy with L2-regularization as the loss function.

$$L = -\frac{1}{n} \sum\nolimits_{i=1}^{N} \sum\nolimits_{k=0}^{C-1} y_{ik} logp\,[k] + \lambda \left\|\theta\right\|_2^2 \tag{27}$$

where n is the number of samples, i is the sample index, k is the class value, λ is the regularization weight, and $\lambda = 10^{-4}$. The optimization algorithm uses the ADAM algorithm [66] based on stochastic gradient descent (SGD). Its parameters are obtained by adaptive learning, and the learning rate is 0.001.

IV. EXPERIMENTS

In this section, experiments are carried out to demonstrate the effectiveness of our model. First, in Section 4.1, the two datasets on which the experiments are conducted are introduced. Then, Section 4.2 presents the evaluation metrics, and baseline methods are described. Section 4.3 discusses the parameters and experimental environments of the inner-class mapping model. Sections 4.4 and 4.5 present the experimental results and the discussion. Finally, four cases under different circumstances are studied in Section 4.6

A. DATASET

1) GETTY IMAGES

To obtain a fine-tuned CNN, a large number of labeled images are needed. The main reason to use Getty Images is that the dataset is already labeled and contains images, labels and relatively formal image descriptions.

While on social media sites, different people may have different descriptions of the same objects, which makes it harder to have a well-labeled training dataset. To improve the accuracy and robustness of our model, we propose using weakly labeled data to train our model in our implementation. We use a list of classes for both sentimental predictions. The list consists of 368 classes and their polarity. Then, query results of images and texts on the Getty Images website are obtained to construct an experimental dataset in line with the classes, and the final weakly labeled dataset that contains 10,496 images and texts is obtained.

2) TWITTER 1269

The Twitter1269 dataset is an open access dataset proposed in [71]. This dataset is a popular image sentiment benchmark and is composed of 1,269 images collected from Twitter. Each image in the dataset was manually labeled by five Amazon Mechanical Turk (AMT) workers as strongly positive (2), positive (1), negative (−1) or strongly negative (−2). These images were ranked according to the sum of their scores by 5 AMT workers and then divided into three confidence level batches:

High confidence (5 agree) images: contains 882 images, five workers all labeled the same sentiment for an image, of which 581 were positive, and 301 were negative.

Midconfidence (4 agree) images: contains 1,116 images labeled by at least four staff with the same sentiment, of which 689 were labeled as positive and 427 as negative.

Low confidence images (3 agree): contains 1,269 images labeled by at least three staff with the same sentiment, of which 769 were labeled as positive and 500 as negative.

B. EVALUATION METRICS AND BASELINES

There are four main evaluation protocols widely used in image sentiment analysis: precision (pre), recall (rec), F-measure (F1) and accuracy rate (Acc). The following open source baselines are used to compare with this model for performance evaluation:

Single textual model: A single textual model is a sentiment analysis method based on textual features. Tan *et al.* [67] proposed a model using multikernel learning to extract text features as the input of a support vector machine to analyze sentiment polarity. Le and Mikolov [68] proposed an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts.

Single visual model: A single visual model refers to a sentiment analysis model based on visual features. Siersdorfer *et al.* [69] proposed a model using low-level visual features extracted by a global color histogram (GCH) for sentiment classification. You *et al.* [37] proposed a progressive CNN (PCNN) method, which uses CNN to extract visual features for regression.

Multimodal model: The multimodal model in this paper refers to a sentiment analysis method consisting of more than one model to extract the feature vector from different datasets. Then, the final polarity is determined by the voting mechanism. Borth *et al.* [51] proposed the Sentibank method. Sentibank is a method carried out on VSO constructed by extracting the ANPs from relevant descriptions. Yuan *et al.* [70] proposed Sentribute to predict image sentiment polarity by using middle-level attributes combined with voting.

Cross-modal model: The cross-modal model uses textual features as a supplement to image sentiment analysis. You *et al.* [71] proposed a transfer learning model based on information entropy to label images. Our proposed cross-modal model uses the MMD to label images and an extra attention-based LSTM to classify sentiment polarity.

C. INNER-CLASS MAPPING PERFORMANCE

We test the inner-class mapping model in this part. First, the image feature vectors and labels are obtained by pretrained CNN and GloVe, respectively. Then, the images in the training set are used as the source domain, and the minimum MMD distance is calculated with the corresponding labels.

In the validation stage, the mapping performance is verified with cross-verification. The experimental results are the average of experiments performed 10 times, and the dataset is shuffled before each experiment. The Getty Images dataset is randomly divided into two partitions of 80% and 20%, and 80% partitions are used as the training set and 20% partition test set.

Experiments are carried out on a workstation with Ubuntu16.04(X86_64) and a NVIDIA GTX1060 GPU. Specifically, the CNN model is initialized using a pretrained 16-layer VGGNet, which includes 13 convolutional layers and 3 fully connected layers, on the Getty Images dataset

TABLE 3. Performance of the inner-class mapping model.

Dataset	Precision	Recall	F1	Accuracy
Getty Images	73.0%	67.5%	70.0%	69.1%

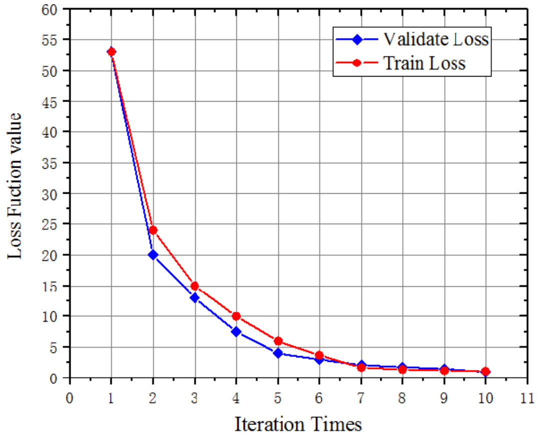


FIGURE 3. Performance of the loss function on the training dataset and validation dataset.

to extract image features. The feature maps are from the conv1_1 layer and the conv4_3 layer, and the output of the conv5_3 layer is used as the input to the inner-class mapping model. The learning rates of the convolutional layers and the last fully connected layer on the classification branch are initialized as 0.001 and 0.01, respectively. All parameters of visual and textual components can be jointly optimized. Experientially, unsupervised learning is used in training with $\Theta = 1$ [63], and the number of training iterations is 10,000 with the GPU. The performance of the inner-class mapping model is shown in Table 3.

D. PERFORMANCE ON GETTY IMAGES

In experiments, the performance of the proposed model is verified with cross-verification. The experimental results are the average of 10 experiments, and the dataset is shuffled before each experiment. The Getty Images dataset is randomly divided into two partitions by 80% and 20%, and 80% partitions are used as the training set and 20% partition test set.

Before validating the proposed model on Getty Images, we saved image descriptions as a file for preprocessing. Preprocessing [72] includes three steps: 1) numbers and special characters in the description are removed; 2) the description file is tokenized with the tokenizer NLTK; 3) words that appear fewer than 5 times are removed and the dimensions of textual feature vectors are limited to 300. Then, the obtained images are divided into several batches for training and testing. In each batch, the parameters of visual and textual components can be jointly optimized. To balance the memory load and convergence rate, each batch size is set to 1,000, and the learning rate is set to 0.01.

TABLE 4. Performance of different models on the Getty Images datasets. Bold represents the best performance of all models.

Methods	Precision	Recall	F1	Accuracy
Tan[68]	71.8%	68.5%	70.1%	65.9%
Mikolov[69]	82.0%	78.8%	80.4%	79.0%
Siersdorfer [70]	68.7%	**84.0%**	75.6%	69.7%
You[37]	73.0%	74.4%	73.7%	71.7%
Borth[51]	74.2%	72.7%	73.4%	67.5%
Yuan [71]	76.9%	69.8%	73.1%	72.7%
You[72]	**84.6%**	75.9%	78.0%	80.0%
Ours	83.2%	79.1%	**81.0%**	**80.6%**

Figure 3 illustrates the change in variance in the loss function as the number of batch iterations increases. The results obtained from the preliminary analysis are that the variation in the loss function and the number of iterations is inversely proportional on randomly selected batches. The loss function converges after approximately 10 iterations.

As shown in Table 4, the recall of single text feature models [67], [68] is generally lower than that of other models, and the recall of the single visual model Siersdorfer *et al.* [69] proposed is the highest (84.0%). The cross-modal model You *et al.* [71] proposed has the highest precision of 84.6%. The precision of our model is similar to that of You *et al.* [71]. The recall of our model is the second highest and is approximately 5% lower than that of the single visual model Siersdorfer *et al.* [69], but F1 reaches 81.0%. The F1 and accuracy of our model are both better than those of other baselines. The cross-modal model You *et al.* [71] carried out image annotation based on information entropy according to labeled images for sentiment analysis. In this paper, the attention-based LSTM is used to obtain the sentiment word to analyze image sentiment. Hence, compared with You *et al.* [71], the precision of our model is approximately 1.4% lower than that of You *et al.* [71], but the advantage of our model is that it can be applied to decompose complex sentences in image descriptions to obtain the image sentiment polarity.

E. PERFORMANCE ON TWITTER1269

In this section, the model proposed in this paper is validated on three batches of Twitter1269 images, and the performance is compared with baselines where the experiments were carried out on the same dataset. Figure 4 illustrates the experimental data on Twitter1269.

As seen in Table 5, the precision, recall, F1, and accuracy of all models decrease with the decline in image confidence, and it is apparent from this table that the subjectivity of different individuals has a great influence on the judgment of image sentiment. However, even if the subjectivity is considered, our model maintains a considerable advantage in precision, F1 value and accuracy. The results of the correlational analysis show that all the evaluation protocols of our model are better than those of the single text feature model. Even though the recall of two single visual models, Siersdorfer *et al.* [69] and You *et al.* [37], are both very high, the overall

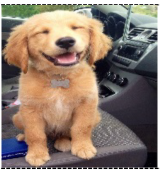
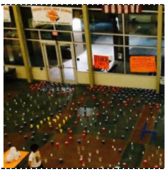
Image						
Caption	Hello there sweetie. :)	This kinda makes me more emotional than sad.	It's near its dying time.	No trading, no killing.	Self-timer will make me more perfect.	I'm still upset over the game!
Index	1	2	3	4	5	6

FIGURE 4. Random samples.

TABLE 5. Performance on Twitter1269 dataset. Bold represents the best performance of all models.

Methods	5 agree				At least 4 agree				At least 3 agree			
	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc
Tan[68]												
Mikolov[69]	74.6%	69.3%	72.7%	72.2%								
Siersdorfer [70]	70.8%	**88.8%**	78.7%	68.4%	68.7%	84.0%	75.6%	66.5%	68.7%	**83.6%**	74.9%	66.0%
You[37]	79.7%	81.1%	83.6%	73.3%	78.6%	**84.2%**	81.1%	75.9%	75.5%	80.5%	77.8%	72.3%
Borth[51]	78.5%	76.8%	77.6%	70.9%	74.2%	72.7%	73.4%	67.5%	72.0%	72.3%	72.1%	66.2%
Yuan[71]	78.9%	82.3%	80.5%	73.8%	75.0%	79.2%	77.1%	70.9%	73.3%	78.3%	75.7%	69.6%
You[72]	83.1%	80.5%	81.8%	80.9%								
Ours	**88.0%**	86.6%	**86.2%**	**82.3%**	**83.1%**	81.2%	**82.1%**	**77.1%**	**79.3%**	77.6%	**78.8%**	**74.2%**

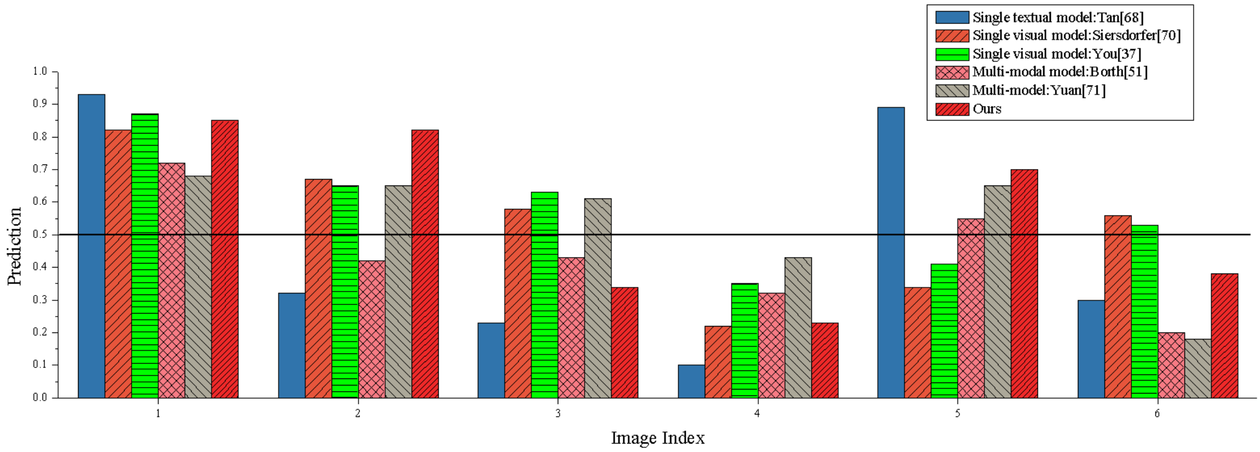


FIGURE 5. Results on 6 random images.

classification precision is obviously lower than that of our model. The results from You *et al.* [37] can be compared with the data from our model, which shows that the precision, recall and F1 are approximately 9, 6 and 1.9% lower, respectively. These results suggest that our model uses the visual and textual features jointly to classify the sentiment polarity together and has obvious advantages.

F. CASE STUDY

This section indicates how the proposed model works through some cases.

1) CASE 1: SENTIMENT ANALYSIS ON RANDOM IMAGES

Images in the Twitter1269 dataset were labeled by AMT staff with negative sentiment as '0' and positive sentiment as '1'. The final image polarity was determined by the predicted probability of IDLSTM. To evaluate the performance of our model directly, six images with different confidence levels are selected as samples with indexes of 1 to 6, and the image descriptions are manually added, as shown in Figure 4.

In Figure 4, images 1, 2, and 5 are positive samples with high confidence levels, image 4 is a negative sample with high confidence levels, and images 3 and 6 are negative samples with low confidence levels. Because the sentiment polarity of the low confidence samples is subjective and uncertain, the high confidence samples are chosen in this paper. Y=0.5 is set as the reference line in the bar chart. The experimental results are shown in Figure 5.

The bar chart in Figure 5 shows the experimental prediction of different models to samples. It is apparent that a single textual model only considers text characters. The textual descriptions of images 2, 3, 4, and 6 are relatively negative, and the predictions of single textual models are all negative, whereas image 2 is actually positive. This means that the single textual model has limitations in social media analysis.

For a single visual model, only visual features are considered; images1, 2, 3, and 6 with distinct colors are identified as positive samples, and images 4 and 5 with gloomy colors are predicted as negative samples. It is obvious that the prediction of image 5 is incorrect. Multimodal model Borth *et al.* [51] predicted image 2 as a negative image, whereas the ground truth of image 2 is positive. A possible explanation for this result might be that Borth *et al.* [51] relies deeply on ANPs. If there are no ANPs in the description or the description does not follow normal syntax, the result of Sentibank will be biased. The multimodal model Yuan *et al.* [70] gives a positive polarity to image 3, whereas image 3 is actually a negative sample. What is notable in Yuan's model [70] is that the essence of the Sentribute method is to train 102 classifiers to distinguish image content attributes. Each classifier is related to one kind of sentiment image scene label; the final result is obtained by voting. However, with a small training size, caution must be applied, as the Sentribute might not be able to recognize all attributes. With the model proposed in this paper, the results of images 1 and 2 are positive, and those of images 3, 4, and 6 are negative. All predictions are correct. In addition, for image 5 with gloomy visual color and a positive textual description, the predicted result is a probability close to 0.5.

However, the accuracy of the inner-class mapping model is not high enough for complex image annotation. The results of this case indicate that the accuracy does not affect the correctness. In particular, this case confirms that our model is suitable for social media sentiment analysis. Further work is required to improve the accuracy of the inner-class mapping model.

2) CASE 2: VISUALIZATION OF ATTENTION

In this section, case 2 is designed to show how the attention-based IDLSTM works by visualization [73]. For this purpose, the output of IDLSTM in each iteration is captured.

Because words in the sentence might belong to the same class, attention weights are appropriate for finding the words related to image labels in the sentence.

As shown in Figure 6, an image randomly selected with the caption ''coffee is better than tea''. ''coffee'' and ''tea'' in the sentence belong to the same class, in which the sentiment ''coffee'' comes from ''better''. When ''coffee'' and ''tea'' with the same class coexist, it is necessary to use the image context as a priori knowledge. Our model repeatedly compares the ''coffee'' linked to the image with other words in the sentence by the attention-based IDLSTM and finally highlights sentimental words as the classification result.

Figure 6 shows the visual results of changing weight in the IDLSTM classification stage. For the attention weight, the attention weight is reflected by the background color of the words. The deeper the color is, the more important the word is. Before the iteration begins (iteration 0), the weights of every word in the image description are distributed equally. With the incremental iterations, the attention weight changes obviously. In the third iteration (iteration 3), ''coffee'' and ''tea'' are highlighted, indicating that class is effectively introduced. In the 10th iteration (iteration 10), the highest weight is assigned to the word ''better'', indicating that weight is changing at the word level, and attention-based IDLSTM can dynamically highlight sentimental words of the entire sentence to make the correct classification.

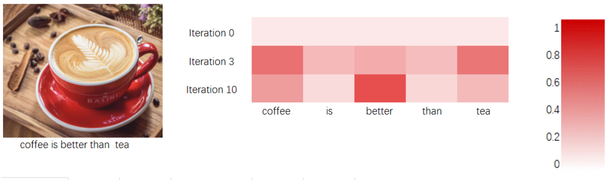


FIGURE 6. Attention weights in IDLSTM.

TABLE 6. Performance of ablation experiments.

Model	Precision	Recall	F1	Accuracy
CNN+IMM	79.6%	**80.3%**	77.1%	75.5%
IDLSTM	78.0%	80.0%	78.9%	79.3%
Whole model	**83.2%**	79.1%	**81.0%**	**80.6%**

3) CASE 3: ABLATION EXPERIMENTS

In this case, ablation experiments are carried out to quantify the effectiveness of the inner-class mapping model (IMM) and inner-class dependency LSTM (IDLSTM) introduced in this paper. The proposed model is retrained by ablating the following components on the Getty Image dataset:

(1) Visual sentiment classifier, where only the image feature is considered. To further study the effect caused by visual features, the IDLSTM is ablated. The visual sentiment classifier (CNN+IMM) consists of CNN and IMM. The outputs of the visual sentiment classifier are image labels, and the polarity of an image is finally calculated by summing the polarity of each image label.

(2) Textual sentiment classifier, where only the textual feature is considered. To study the effect caused by IDLSTM, CNN and IMM are ablated. The output of IDLSTM is the polarity of the description.

Because of the uncertainty of Twitter1269, three models are tested to quantify the effectiveness on Getty Images. Table 6 illustrates the performance of ablation experiments. Compared with CNN+IMM, it is obvious that IDLSTM can improve the performance of the model. The recall of the whole model is lower by more than 1%, but the precision, F1 and accuracy are higher by 3.2, 4, and 5%, respectively. Compared with the single IDLSTM, the performance of the whole model is better than that of the single IDLSTM. The precision, F1 and accuracy are higher by 5%, 2%, and 1%, respectively. This means that the inner-class mapping model is useful.

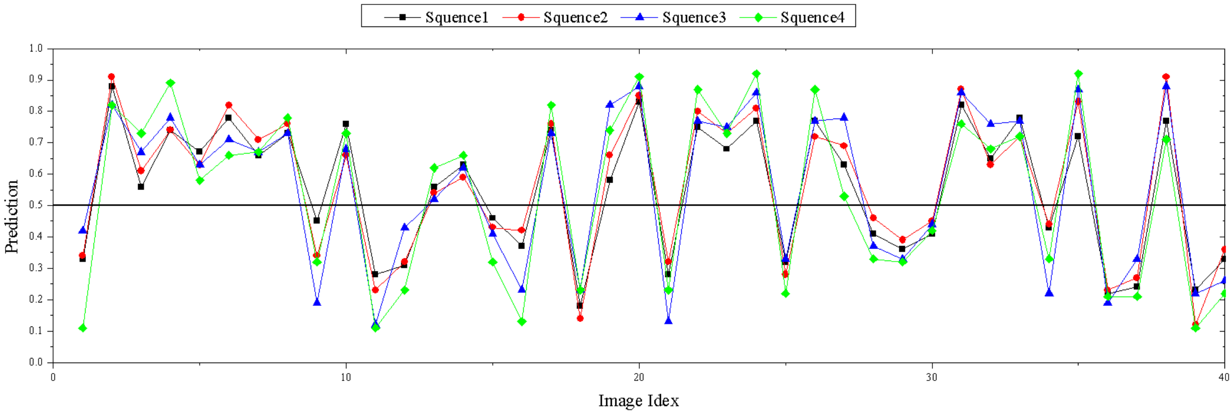


FIGURE 7. Result on 40 random images.

Consequently, it can be concluded that there exists inner-class dependency, and the exploitation of the correlations between image and description is conducive to more effective cross-modal image sentiment classification.

4) CASE 4: THE INFLUENCE OF DIFFERENT INPUT SEQUENCES

LSTM carries out sentiment analysis based on memory slots that are influenced by the input sequence. In fact, the output of LSTM is affected by the input sequence.

In this experiment, 40 images from the Twitter1269 dataset with high confidence levels were randomly selected as the samples, index 1 to 40, and image descriptions were also manually added. The class [41] order was sorted randomly, and four sorted queues with different orders were taken as inputs of IDLSTM. In Figure 7, the horizontal axis represents the image index, and the vertical axis is the positive probability of an image. Figure 7 illustrates the positive probability of each image predicted.

As seen in Figure 7, the four output curves are different according to the four different input class sequences. Although the differences in the input class sequences leads to a certain deviation in the predicted probability, there is only a small-scale fluctuation in each curve and no instances in which the polarity prediction is upside down on the reference line due to the difference in the input sequence of IDLSTM.

It is shown that the output of the proposed model is independent of the input class sequence of IDLSTM.

V. CONCLUSION

In this paper, a joint visual-textual cross-modal sentiment analysis model is proposed. This model extracts visual object features and uses them as the attention weight parameter of LSTM to obtain the context image object related in the corresponding textual description. The sentiment polarity of an image is then obtained. This model can not only solve a multiobjective image sentiment analysis but also improve the utilization rate of semantic correlation descriptions. In experiments, the Getty Images and Twitter1269 datasets are used to validate the proposed sentiment analysis model. The results show that the proposed model outperforms existing state-of-the-art models on social media image datasets.

However, there are still some unsatisfactory problems in the operation of the model in experiments, such as memory overhead, long system runtime, and limitations in some special application scenarios. Future research should be undertaken to investigate the following. 1) Improvement of the precision of the inner-class mapping model on transfer learning. For example, using a knowledge graph to provide prior knowledge for target text feature mapping. 2) Model parameter optimization and structure reconstruction. 3) The application to other domains, such as audio-video domain adaptation.

REFERENCES

[1] E. Cambria, D. Das, and S. Bandyopadhyay, “Affective computing and sentiment analysis,” in *A Practical Guide to Sentiment Analysis*. Springer, 2017, ch. 1, pp. 1–15.

[2] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, “Chinese text sentiment analysis based on extended sentiment dictionary,” *IEEE Access*, vol. 7, pp. 43749–43762, 2019.

[3] A. Bandhakavi, N. Wiratunga, S. Massie, and D. Padmanabhan, “Lexicon generation for emotion detection from text,” *IEEE Intell. Syst.*, vol. 32, no. 1, pp. 102–108, Jan. 2017.

[4] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, “Sentiment analysis leveraging emotions and word embeddings,” *Expert Syst. Appl.*, vol. 69, pp. 214–224, Mar. 2017.

[5] F. G. Contratres, S. N. Alves-Souza, and L. V. L. Filgueiras, “Sentiment analysis of social network data for cold-start relief in recommender systems,” in *Proc. World Conf. Inf. Syst. Technol.* Cham, Switzerland: Springer, 2018, pp. 122–132.

[6] V. Rozgic, S. Ananthakrishnan, and S. Saleem, “Ensemble of SVM trees for multimodal emotion recognition,” in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.

[7] V. Singh, M. Ram, and B. Pant, “Identification of zonal-wise passenger’s issues in Indian railways using latent Dirichlet allocation (LDA): A sentiment analysis approach on tweets,” *Math. Appl. Inf. Syst.*, vol. 2, no. 1, pp. 265–276, 2018.

[8] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, and E. Cambria, “Bayesian network based extreme learning machine for subjectivity detection,” *J. Franklin Inst.*, vol. 355, no. 4, pp. 1780–1797, Mar. 2018.

[9] E. Daglarli, H. Temeltas, and M. Yesiloglu, “Behavioral task processing for cognitive robots using artificial emotions,” *Neurocomputing*, vol. 72, nos. 13–15, pp. 2835–2844, Aug. 2009.

[10] E. Cambria, S. Poria, and D. Hazarika, "SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1795–1802.

[11] M. Dragoni, S. Poria, and E. Cambria, "OntoSenticNet: A commonsense ontology for sentiment analysis," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 77–85, May 2018.

[12] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D. Z. Rodriguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Trans Ind. Informat.*, vol. 15, no. 4, pp. 2124–2135, Apr. 2019.

[13] P. V. Kulkarni, M. B. Nagori, and V. P. Kshirsagar, "An in-depth survey of techniques employed in construction of emotional lexicon," in *Information and Communication Technology for Intelligent Systems*. Singapore: Springer, 2019, pp. 609–620.

[14] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018.

[15] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[16] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A generative model for category text generation," *Inf. Sci.*, vol. 450, pp. 301–315, Jun. 2018.

[17] X. Yang, S. Xu, H. Wu, and R. Bie, "Sentiment analysis of weibo comment texts based on extended vocabulary and convolutional neural network," *Procedia Comput. Sci.*, vol. 147, pp. 361–368, Jan. 2019.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[19] S. Xiong, K. Wang, D. Ji, and B. Wang, "A short text sentiment-topic model for product reviews," *Neurocomputing*, vol. 297, pp. 94–102, Jul. 2018.

[20] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 45–63, Oct. 2014.

[21] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on Twitter aspect-based sentiment analysis," *Int. J. Speech Technol.*, pp. 1218–1232, Dec. 2017.

[22] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[23] G. Liu, X. Huang, X. Liu, and A. Yang, "A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network," *Comput. J.*, May 2019, doi: 10.1093/comjnl/bxz031.

[24] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 5876–5883.

[25] D. Tang, B. Qin, and X. Feng, "Effective LSTMs for target-dependent sentiment classification," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 3298–3307.

[26] K. Lu and J. Wu, "Sentiment analysis of film review texts based on sentiment dictionary and SVM," in *Proc. 3rd Int. Conf. Innov. Artif. Intell. (ICIAI)*, 2019, pp. 73–77.

[27] L. Bin, L. Quan, X. Jin, Z. Qian, and Z. Peng, "Aspect-based sentiment analysis based on multi-attention CNN," *J. Comput. Res. Development.*, vol. 54, no. 8, pp. 1724–1735, 2017.

[28] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[29] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Inf. Fusion*, vol. 44, pp. 65–77, Nov. 2018.

[30] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, 2012, pp. 229–238.

[31] S. Wang, J. Wang, Z. Wang, and Q. Ji, "Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2185–2197, Dec. 2015.

[32] D. Borth, T. Chen, and R. Ji, "SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 459–460.

[33] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[34] J. Wang, J. Fu, and Y. Xu, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3484–3490.

[35] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, Sep. 2017.

[36] J. K. Rout, K.-K.-R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," *Electron. Commerce Res.*, vol. 18, no. 1, pp. 181–199, Apr. 2017.

[37] Q. You, J. Luo, and H. Jin, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 39th AAAI Conf. Artif. Intell. (AAAI)*, 2015, pp. 381–388.

[38] U. Ahsan, M. De Choudhury, and I. Essa, "Towards using visual attributes to infer image sentiment of social events," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1372–1379.

[39] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, Oct. 2018.

[40] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[41] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[42] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowl.-Based Syst.*, vol. 167, pp. 26–37, Mar. 2019.

[43] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, Jul. 2019.

[44] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "YouTube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May 2013.

[45] S. Kumar, M. Yadava, and P. P. Roy, "Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction," *Inf. Fusion*, vol. 52, pp. 41–52, Dec. 2019.

[46] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[47] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, pp. 217–230, Oct. 2017.

[48] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[49] S. P. Byrne, A. Mayo, C. O'Hair, M. Zankman, G. M. Austin, C. Thompson-Booth, E. J. McCrory, L. C. Mayes, and H. J. V. Rutherford, "Facial emotion recognition during pregnancy: Examining the effects of facial age and affect," *Infant Behav. Develop.*, vol. 54, pp. 108–113, Feb. 2019.

[50] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, and Y. He, "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Appl. Soft Comput.*, vol. 80, pp. 387–399, Jul. 2019.

[51] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 223–232.

[52] C. G. Maurya, S. Gore, and D. S. Rajput, "A use of social media for opinion mining: An overview (with the use of hybrid textual and visual sentiment ontology)," in *Proc. Int. Conf. Recent Adv. Comput. Commun.* Singapore: Springer, 2018, pp. 315–324.

[53] Z. Li, Y. Fan, W. Liu, and F. Wang, "Image sentiment prediction based on textual descriptions with adjective noun pairs," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1115–1132, Jan. 2017.

[54] J. Teng-Jiao, W. Chang-Xuan, L. De-Xi, L. Xi-Ping, and L. Guo-Qiong, "Extracting Target-opinion pairs based on semantic analysis," *Chin. J. Comput.*, vol. 40, no. 3, pp. 617–633, 2017.

[55] Y.-H.-H. Tsai, Y.-R. Yeh, and Y.-C.-F. Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5081–5090.

[56] X. Huang, Y. Peng, and M. Yuan, "Cross-modal common representation learning by hybrid transfer network," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1893–1900.

[57] P. Schmitter, J. Steinrücken, C. Römer, A. Ballvora, J. Léon, U. Rascher, and L. Plümer, ''Unsupervised domain adaptation for early detection of drought stress in hyperspectral images,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 131, pp. 65–76, Sep. 2017.

[58] R. Ji, F. Chen, L. Cao, and Y. Gao, ''Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning,'' *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.

[59] A. Van Opbroek, H. C. Achterberg, M. W. Vernooij, and M. De Bruijne, ''Transfer learning for image segmentation by combining image weighting and kernel learning,'' *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 213–224, Jan. 2019.

[60] F. Wu, Z. Wang, Z. Zhang, Y. Yang, J. Luo, W. Zhu, and Y. Zhuang, ''Weakly semi-supervised deep learning for multi-label image annotation,'' *IEEE Trans. Big Data*, vol. 1, no. 3, pp. 109–122, Sep. 2015.

[61] W. Wang, J. Shen, and H. Ling, ''A deep network solution for attention and aesthetics aware photo cropping,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.

[62] W. Wang, J. Shen, and L. Shao, ''Video salient object detection via fully convolutional networks,'' *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[63] J. Wang, Y. Chen, L. Hu, X. Peng, and P. S. Yu, ''Stratified transfer learning for cross-domain activity recognition,'' in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. (PerCom)*, Mar. 2018, pp. 1–10.

[64] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, ''A kernel two-sample test,'' *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[65] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, ''Domain adaptation via transfer component analysis,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[66] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15.

[67] J. Tan, M. Xu, and L. Shang, ''Sentiment analysis for images on microblogging by integrating textual information with multiple kernel learning,'' in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2016, pp. 496–506.

[68] Q. V. Le and T. Mikolov, ''Distributed representations of sentences and documents,'' in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

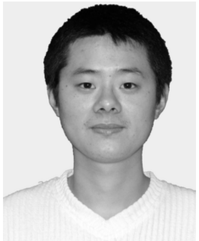
[69] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, ''Analyzing and predicting sentiment of images on the social Web,'' in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 715–718.

[70] J. Yuan, S. Mcdonough, Q. You, and J. Luo, ''Sentribute: Image sentiment analysis from a mid-level perspective,'' in *Proc. 2nd Int. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, 2013, pp. 1–8, doi: 10.1145/2502069.2502079.

[71] Q. You, J. Luo, H. Jin, and J. Yang, ''Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia,'' in *Proc. 9th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2016, pp. 13–22.

[72] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann, ''Web and personal image annotation by mining label correlation with relaxed visual graph embedding,'' *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[73] K. Xu, J. Ba, and R. Kiros, ''Show, attend and tell: Neural image caption generation with visual attention,'' in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.



KE ZHANG was born in Jiangsu, China, in 1989. He received the B.S. degree from the Jinling College, Nanjing University, in 2012, and the M.S. degree from the Guilin University of Electronic Technology. He is currently pursuing the Ph.D. degree with Shanghai University. His research areas are in interdisciplinary areas on image recognition and natural language processing.

YUNWEN ZHU was born in Xiamen, Fujian, China, in 1995. She received the B.S. degree from Shanghai University, in 2016, where she is currently pursuing the Ph.D. degree with the Shanghai File Academy. She is interested in interdisciplinary research, such as fuzzy theory and deep learning.

WENJUN ZHANG was born in Shanghai, in November 1959. He received the Dipl.Ing., master's, and Ph.D. degrees from the University of Belgrade, in 1984, 1986, and 1989, respectively. He was given an accelerated promotion to an Associate Professor in 1991, and a Professor in 1995. From 1999 to 2003, he was a Diplomatic Officer with the Science and Technology Cooperation, Chinese Embassy, Belgrade. He is currently an Academic Leader of the Digital Media and Innovation Engineering, Shanghai University. He is also the Dean of the School of Information Technology, Shanghai Jianqiao University. His research interests include digital media technology and applications, digital image processing, digital content design and productions, and computational electromagnetics.

WEILIN ZHANG received the B.S. degree in computer science from Shanghai University, Shanghai, China, in 2018, where he is currently pursuing the M.S. degree. His research areas are search engine, recommendation systems, and machine learning.

YONGHUA ZHU was born in Zhejiang, China, in 1967. He received the B.S. degree from the Department of Information and Control Engineering, Xi'an Jiaotong University, the M.S. degree from the Department of Electrical Engineering, Shanghai Tongji University, and the Ph.D. degree from the School of Communication and Information Engineering, Shanghai University. Since 2010, he has been a Teacher with the School of Computer Engineer and Science, Shanghai University, where he is currently an Associate Professor and a Supervisor with Shanghai Film Academy. His research interests include software engineering and machine learning, especially in interdisciplinary sentiment analysis.