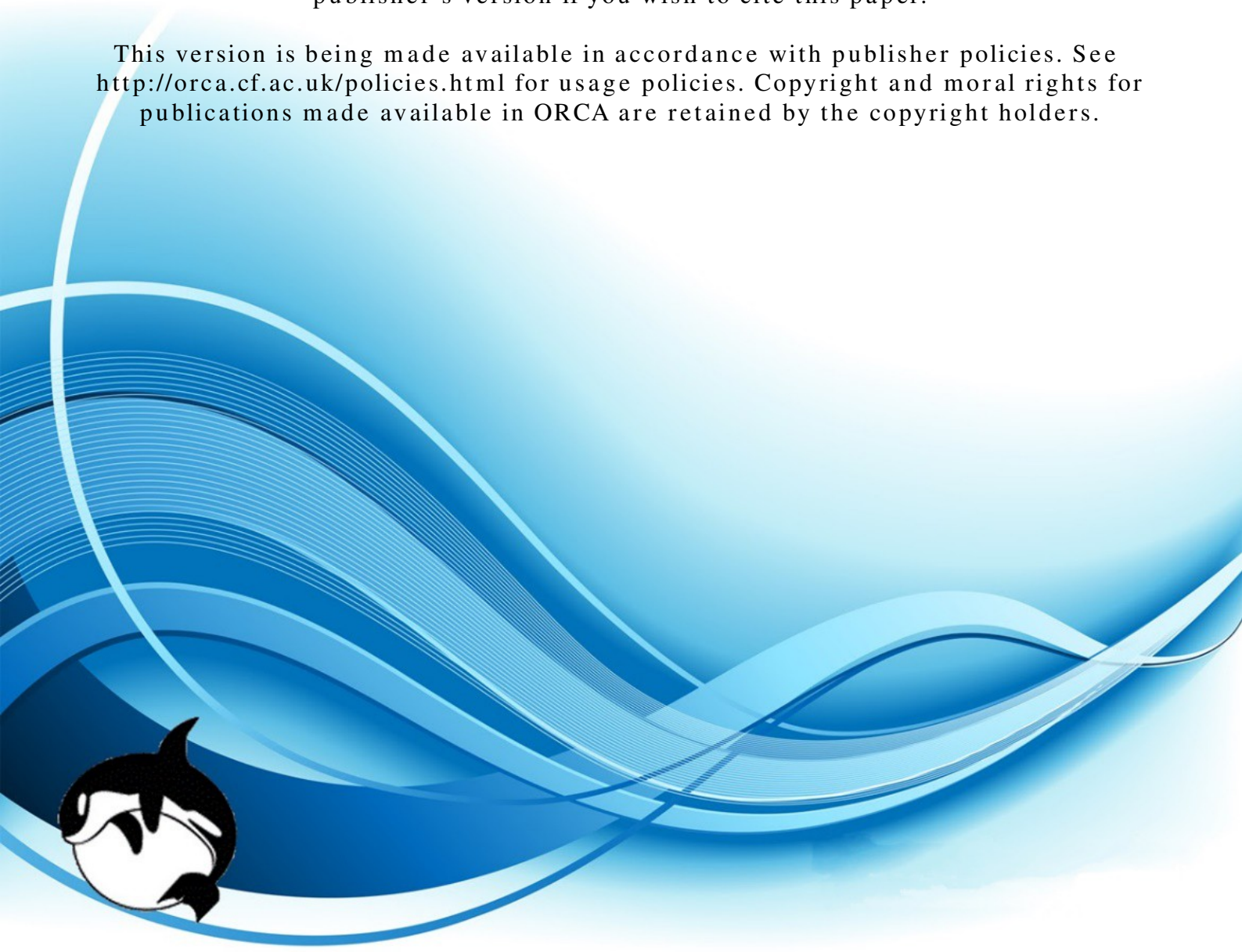


Multimodal Sentiment Analysis With Image-Text Interaction Network

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian

***Abstract*—More and more users are getting used to posting images and text on social networks to share their emotions or opinions. Accordingly, multimodal sentiment analysis has become a research topic of increasing interest in recent years. Typically, there exist affective regions that evoke human sentiment in an image, which are usually manifested by corresponding words in people's comments. Similarly, people also tend to portray the affective regions of an image when composing image descriptions. As a result, the relationship between image affective regions and the associated text is of great significance for multimodal sentiment analysis. However, most of the existing multimodal sentiment analysis approaches simply concatenate features from image and text, which could not fully explore the interaction between them, leading to suboptimal results. Motivated by this observation, we propose a new image-text interaction network (ITIN) to investigate the relationship between affective image regions and text for multimodal sentiment analysis. Specifically, we introduce a cross-modal alignment module to capture region-word correspondence, based on which multimodal features are fused through an adaptive cross-modal gating module. Moreover, considering the complementary role of context information on sentiment analysis, we integrate the individual-modal contextual feature representations for achieving more reliable prediction. Extensive experimental results and comparisons on public datasets demonstrate that the proposed model is superior to the state-of-the-art methods.**

***Index Terms*—Multimodal sentiment analysis, Image-text interaction, Region-word alignment.**

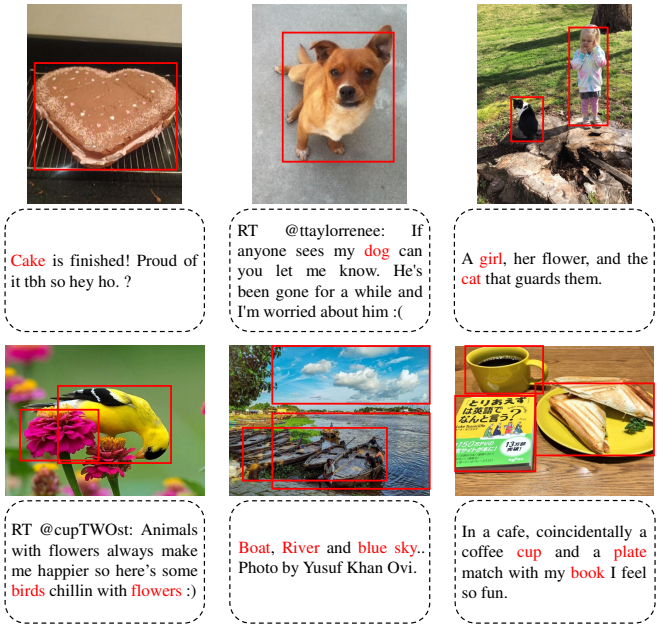


Fig. 1. Examples of image-text pairs from Twitter. The sentence words, like "cake", "girl", and "boat", highlighted in red, are manifested by regions of the corresponding images. Human sentiment can be evoked mostly by the affective regions in an image.

I. Introduction

With the booming of mobile devices and social networks, people tend to post images and text together to express their emotions or opinions. Aiming at analyzing and recognizing sentiments from data of diverse modalities, multimodal sentiment analysis has attracted increasing research attention in recent years [1]. Understanding multimodal sentiment has various potential applications, including personalized advertising [2], affective cross-modal retrieval [3], opinion mining [4], decision making [5] and so on.

In contrast to single-modal sentiment analysis, processing and analyzing data from different modalities bring both opportunities and challenges. Early multimodal works are mainly based on handcrafted features. However, handcrafted features are typically designed with limited human knowledge and are unable to describe the high abstraction of sentiment comprehensively, which leads to suboptimal results. In recent years, with the development of deep learning, convolutional neural networks have been utilized in multimodal sentiment analysis with encouraging performance. However, most of them either simply concatenate features extracted from different modalities [6], or learn the relation between image and text at a coarse level [7]. Intuitively, human sentiment can be evoked mostly by affective regions in an image [8], [9], [10], which are usually manifested on some words of the associated comment text. In the same way, when people compose comment of an image, they often describe the affective regions of the

This work was supported by the National Natural Science Foundation of China under Grants 62171340, 61771473 and 61991451, the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024, and the Six Talent Peaks High-level Talents in Jiangsu Province under Grant XYDXX-063 (*Corresponding author: Leida Li*).

T. Zhu is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: zhutong@cumt.edu.cn).

L. Li is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: ldli@xidian.edu.cn).

J. Yang is with the School of Computer Science and Control Engineering, Nankai University, Tianjin 300350, China (e-mail: yangjufeng@nankai.edu.cn).

S. Zhao is with the Department of Radiology, Columbia University, USA (e-mail: schzhao@gmail.com).

H. Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. (e-mail: hantao.liu@cs.cardiff.ac.uk).

J. Qian is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: qianjsh@cumt.edu.cn).

image subconsciously. Some examples collected from Twitter are illustrated in Fig. 1. We naturally pay more attention to the dog in the second image at first glance, and meanwhile, the word "dog" can be found in the corresponding sentence. Similarly, the word "girl" described in the third sentence is also corresponding to a salient region in the image. Likewise, we can find "cup", "plate" and "book" from both image and text in the last example. Therefore, the corresponding relationship between regions and words does exist in image-text pairs, and such relationship can be considered as one type of cross-modal interaction. Taking into account this particular information would be beneficial for multimodal sentiment analysis.

To address the above problem, this paper presents an image-text interaction network for multimodal sentiment analysis, which focuses on the alignment between image regions and text words and integrates both visual and textual context information. The proposed image-text interaction network comprises a cross-modal alignment module and a cross-modal gating module. The cross-modal alignment module selects the word-level textual information for each region with a cross-modal attention mechanism. The cross-modal gating module utilizes a soft gate to fuse multimodal features adaptively, which can eliminate the influence of misaligned region-word pairs and further strengthen the interactions between image and text. Through these two modules, cross-modal interactions can be captured by aligning image regions and sentence words. In addition, considering the complementary role of context information on sentiment analysis, namely the same object or the same word in different context may evoke different emotions, the proposed method also integrates the individual-modal context representations to explore multimodal sentiment more comprehensively.

The main contributions of this paper can be summarized as follows:

- We propose a novel image-text interaction network for multimodal sentiment analysis. The proposed method explicitly aligns affective image regions and text words for analyzing the image-text interaction. Individual modality visual and textual context features are also incorporated to achieve more comprehensive prediction.
- We introduce a cross-modal alignment module based on cross-modal attention mechanism, which captures fine-grained correspondences between image regions and text words. To suppress the negative effect of misaligned region-word pairs, an adaptive cross-modal gating module is also proposed to fuse multimodal features, which further strengthens the deep interactions between image and text.
- We have done extensive experiments and comparisons on public databases to demonstrate the advantages of the proposed method. Ablation studies are also conducted and the results demonstrate the rationality of the proposed approach.

The rest of this paper is organized as follows. Section II summarizes related works on sentiment analysis. The proposed approach is detailed in Section III. Experiments, including comparisons, ablation studies, hyperparameter analysis and visualization, are given in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

In this section, we briefly review the existing methods for sentiment analysis, including single-modal and multimodal approaches.

A. Single-modal Sentiment Analysis

1) Visual Sentiment Analysis: Early studies on visual sentiment analysis mainly concentrated on designing handcrafted features for modeling image emotion. Based on psychology and art theory, Machajdik *et al*. [11] extracted low-level features, *e.g.*, composition, texture and color, to predict image emotions. Zhao *et al*. [12] utilized principles-of-art-based emotion features for image emotion classification and regression, including balance, emphasis, harmony, variety, gradation, and movement. Borth *et al*. [13] introduced SentiBank to detect Adjective Noun Pairs (ANPs) from an image, which can be considered as mid-level features on the basis of visual concepts. Yuan *et al*. [14] presented an image sentiment prediction approach, Sentribute, by leveraging 102 mid-level attributes, making the classification results more interpretable. Zhao *et al*. [15] combined features of different levels with multi-graph learning, including generic and elements-of-art based low-level features, attributes and principles-of-art based mid-level features, and semantic concepts and facial expressions based high-level features.

In recent years, deep neural networks have been widely adopted in visual sentiment analysis. Chen *et al*. [16] utilized deep convolutional neural networks (CNNs) and introduced a visual sentiment concept classification model named DeepSentiBank. You *et al*. [17] proposed a progressive CNN training for predicting image emotions, which makes use of images labeled with website meta data. Taking into consideration that an image usually evokes a mixture of diverse emotions, Yang *et al*. [18] developed a multi-task framework to jointly optimize visual emotion distribution and classification learning. You *et al*. [19] investigated the sentiment-related local image areas through attention mechanism, and train a sentiment classifier based on these local features. Yang *et al.* [8] introduced a weakly supervised coupled convolutional network (WSCNet) to detect the sentiment map, which utilizes both holistic and local features for visual emotion prediction.

2) Text Sentiment Analysis: Text sentiment classification approaches can be categorized into two types, namely lexicon-based models and machine learning models. Hu *et al.* [20] predicted the semantic orientation of opinion sentences by using adjectives as prior positive or negative polarity. Taboada *et al.* [21] proposed a lexicon-based method named Semantic Orientation CALculator (SO-CAL), which not only employs dictionaries of words annotated with semantic orientation but also integrates intensification and negation to analyze text sentiment. Pang *et al.* [22] first applied machine learning for text sentiment classification, including Naive Bayes, support vector machines and maximum entropy classification. To capture semantic as well as sentiment information among words,

Negative (a) *Positive* (b) *Positive* (c)

Fig. 2. Some examples to illustrate the necessity of predicting sentiments using multimodal data. Sentiment cannot be accurately evaluated by a single modality.

Maas *et al.* [23] presented unsupervised probabilistic model of documents to learn semantic similarities, and a supervised model to predict sentiment annotations. Barbosa *et al.* [24] proposed a two-step sentiment classification approach for Twitter messages, which utilizes online labels as the training data.

Motivated by the superior performance of deep learning models on natural language processing (NLP), Kim [25] first utilized CNN for text sentiment classification. Tai *et al.* [26] considered complicated sentence structure and introduced the tree-structured long short-term memory (Tree-LSTM) for sentence sentiment classification. To effectively model document representation, Tang *et al.* [27] first used CNN and LSTM to get sentence representations, and then exploited gated recurrent neural network to encode sentence semantics and their inherent relations. Yang *et al.* [28] developed a hierarchical attention network (HAN) for document-level sentiment classification task, which employs the attention mechanism to assist networks in selecting important words and sentences. Considering that the concerned aspect is closely related to the sentiment polarity of a sentence, Wang *et al.* [29] introduced a LSTM network based on the attention mechanism to improve aspect-level sentiment classification.

Intuitively, sentiment is highly subjective and extremely complex. However, visual and textual sentiment analysis only extract features from single modality information, which is unable to represent sentiment comprehensively. In fact, sentiments expressed in social media are usually in the form of multiple modalities. Therefore, in this paper, we concentrate on dealing with multimodal sentiment analysis.

B. Multimodal Sentiment Analysis

Psychologists and engineers have demonstrated that sentiment is mainly determined by the joint effect of multimodal data [30], [31], [32]. The same image accompanied with different text may evoke opposite sentiments. Some examples are illustrated in Fig. 2. As shown in Fig. 2(a), the smiling girl in the image conveys us positive feeling. However, we are actually evoked with negative sentiment from the corresponding text. Similarly, the man in Fig. 2(c) makes us feel negative, while the image-text pair is trying to express positive sentiment. Furthermore, we cannot figure out specific emotion from the image in Fig. 2(b). The positive feeling can only be inferred when we look at the text. Therefore, a single modality is not sufficient for predicting the accurate sentiment, and it is necessary to analyse sentiments with multimodal data. Multimodal sentiment analysis takes advantage of features from different modalities for overall sentiment prediction. Wang *et al*. [33] proposed a cross-media bag-of-words model (CBM) for Microblog sentiment classification, where text and images are represented as unified bag-of-words. You *et al.* [34] proposed a cross-modality consistent regression (CCR) method, which uses visual and textual features for joint sentiment prediction. Xu *et al*. [6] developed a hierarchical semantic attentional network (HSAN), and image caption was utilized as semantic information to help analyze multimodal sentiment. To fully capture detailed semantic information, Xu *et al*. [35] presented a deep network named MultiSentiNet, which leverages scene and object features of image for pointing out important sentence words based on attention. Considering that information from two different modalities can affect and supplement each other, Xu *et al.* [7] developed a co-memory network modeling the mutual influences between image and text iteratively to classify multimodal sentiment. Zhao *et al.* [36] proposed an image-text consistency driven method, where textual features, social features, low-level and mid-level visual features and image-text similarities are utilized. Poria *et al.* [37] introduced a LSTM-based approach to model the interdependencies and relations among utterances for multimodal sentiment prediction. Truong *et al*. [38] proposed Visual Aspect Attention Network (VistaNet), which incorporates images as attention to help find out sentences that are important for document sentiment classification. Motivated by text-based aspect-level sentiment analysis, Xu *et al*. [39] presented a new task, aspect-based multimodal sentiment analysis, and provided a new public multimodal aspect-level sentiment dataset.

Since multimodal fusion plays a significant role in multimodal sentiment analysis, several works have been proposed with focus on designing fusion strategies among different modalities. Poria *et al.* [40] proposed a fusion network based on attention mechanism (AT-Fusion) to fuse features from different modalities. Zadeh *et al*. [41] developed a new fusion approach termed Tensor Fusion for multimodal sentiment analysis, which models unimodal, bimodal and trimodal dynamics explicitly. Huang *et al.* [42] introduced an image-text sentiment analysis method, namely Deep Multimodal Attentive Fusion (DMAF), which employs intermediate fusion and late fusion for combining the unimomdal features and the internal cross-modal correlation. The Gated Multimodal Unit (GMU) model was proposed by Arevalo *et al.* [43] for data fusion through learning to control how much input modalities influence the unit activation using gates. Liu *et al*. [44] proposed the Low-rank Multimodal Fusion, an approach leveraging low-rank tensors for multimodal fusion, and demonstrated its efficiency on the sentiment analysis task. To filter out conflicting information from different modalities, Majumder *et al*. [45] devised a hierarchical feature fusion scheme that proceeds from bimodal data fusion to trimodal data fusion.

Despite the impressive advances in multimodal sentiment tasks, little attention has been paid in exploring cross-modal

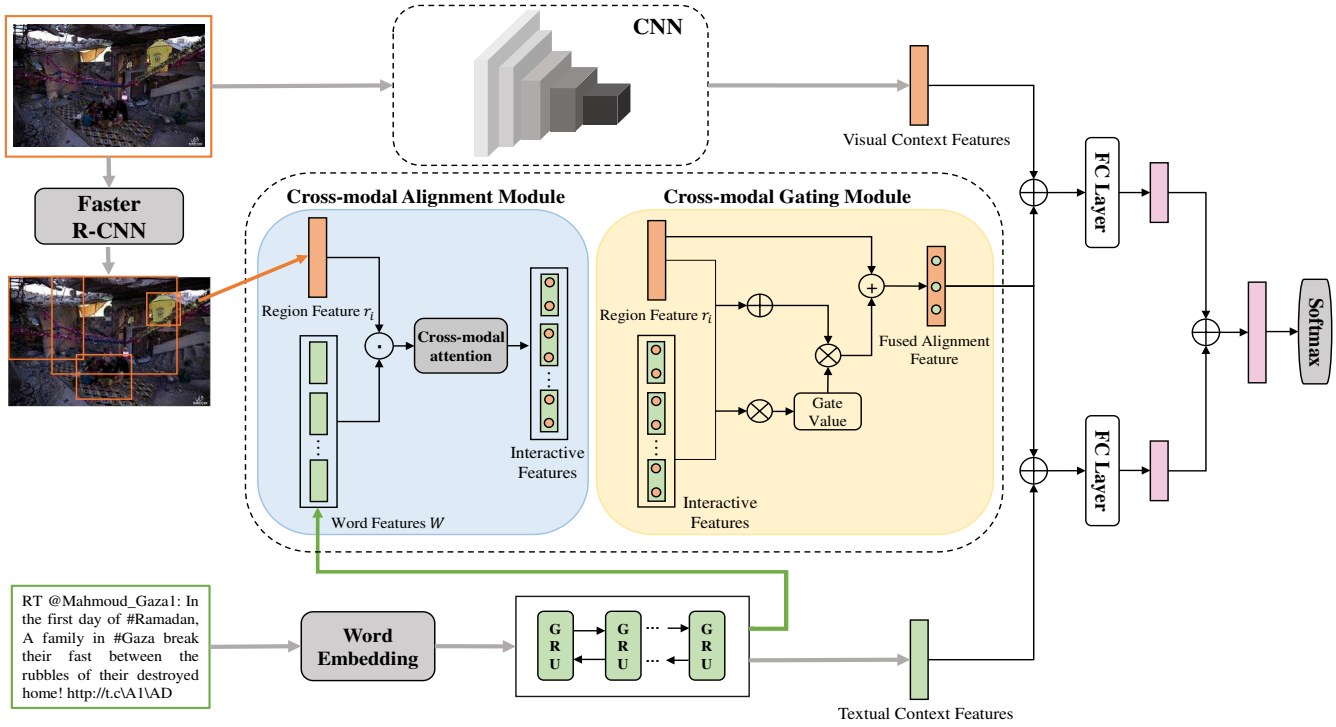


Fig. 3. The overall framework of the proposed Image-Text Interaction Network (ITIN) for multimodal sentiment analysis. The latent alignment between image regions and sentence words is achieved using a Cross-modal Alignment Module and a Cross-modal Gating Module. The visual and textual context representations are further integrated for exploring multimodal sentiment more comprehensively.

interactions for image-text sentiment analysis. Most of the existing approaches simply concatenate features extracted from different modalities or learn the relation between image and text at a coarse level, which leads to suboptimal predictions. Considering the mutual influences and intricate relationship between the two modalities, we propose an image-text interaction network (ITIN) for multimodal sentiment analysis. Specially, we capture the latent alignment between image regions and text words to explore cross-modal interactions at a finer level, which are mainly achieved using the proposed cross-modal alignment module and cross-modal gating module. Besides, contextual features of individual modalities are also integrated into our network considering their complementary roles for sentiment prediction.

III. Proposed Method

In this section, we elaborate on the details of the proposed Image-Text Interaction Network (ITIN) for multimodal sentiment analysis. In contrast to the existing approaches, we focus on the alignment between image regions and text, aiming to investigate the cross-modal interactions at a finer level. Single-modal visual and textual context information are also integrated to achieve more comprehensive prediction. The overall architecture of the proposed model is shown in Fig. 3. Given an input image-text pair, we first extract and encode image regions and sentence words simultaneously. Then we explore the correspondence between local fragments from different modalities with the proposed cross-modal alignment module. An adaptive cross-modal gating module is then introduced to decide how much interactive information should be passed. Furthermore, we incorporate visual and textual context representations with local alignment features respectively. Finally, the multimodal sentiment label is predicted through several multilayer perceptron and a softmax classifier.

A. Image-Text Interaction

1) Cross-modal Alignment Module: For the input image $\boldsymbol{I}$, following [46], we detect image regions and their associated representations utilizing Faster R-CNN [47], which is pre-trained on Visual Genomes dataset [48] using ResNet-101 [49] as backbone. This model predicts object classes as well as attribute classes simultaneously, which can provide richer visual semantic information. We select the top m region proposals for each image. For each region, a 2048-dimensional feature vector is extracted by average-pooling operation, which is denoted as $\boldsymbol{f}_i, i = 1, 2, ..., m$. We then transform $\boldsymbol{f}_i$ to a d-dimentional region feature $\boldsymbol{r}_i$ via a linear projection layer:

$$\boldsymbol{r}_i = \mathbf{W}_r \boldsymbol{f}_i + \boldsymbol{b}_r, i \in [1, m], \tag{1}$$

where $\mathbf{W}_r$ and $\boldsymbol{b}_r$ are learnable parameters, and $\boldsymbol{r}_i$ is the ith region feature.

For an input sentence $\boldsymbol{T}$ with n words, we apply pre-trained BERT-Base [50] to embed each word into a 768-dimensional embedding vector $\boldsymbol{x}_i, i \in [1, n]$. Then, we employ a bidirectional GRU [51], [52] to summarize context information in the sentence, which is achieved by:

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{GRU}(\boldsymbol{x}_i, \overrightarrow{\boldsymbol{h}_{i-1}}), i \in [1, n], \tag{2}$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{GRU}(\boldsymbol{x}_i, \overleftarrow{\boldsymbol{h}_{i+1}}), i \in [1, n], \tag{3}$$

where $\overrightarrow{\boldsymbol{h}_i} \in \mathbb{R}^d$ denotes the forward hidden state and $\overleftarrow{\boldsymbol{h}_i} \in \mathbb{R}^d$ denotes the backward hidden state. The final word feature $\boldsymbol{w}_i$ is defined as the average of bidirectional hidden states:

$$\boldsymbol{w}_i = \frac{\overrightarrow{\boldsymbol{h}_i} + \overleftarrow{\boldsymbol{h}_i}}{2}, i \in [1, n]. \tag{4}$$

Given a region-level feature set $\boldsymbol{R} = \{\boldsymbol{r_1}, ..., \boldsymbol{r_m}\}$ and a word-level feature set $\boldsymbol{W} = \{\boldsymbol{w_1}, ..., \boldsymbol{w_n}\}$, the cross-modal alignment module aims to align image region and sentence word in the embedding space. Based on cross-modal attention mechanism, this module attends on sentence words with respect to each image region, which discovers the most corresponding textual information for each region.

Following [53], the region-word affinity matrix is first computed as:

$$\mathbf{A} = (\hat{\mathbf{W}}_r \boldsymbol{R})(\hat{\mathbf{W}}_t \boldsymbol{W})^T, \tag{5}$$

where $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_t$ represent the projection matrices to obtain k-dimensional region and word features. For the region-word affinity matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A}_{ij}$ indicates the affinity between the ith region and the jth word.

To infer the latent alignments between local fragments from different modalities, we attend on each word with respect to each region by further normalizing the affinity matrix $\mathbf{A}$ as:

$$\bar{\mathbf{A}} = \text{softmax}\left(\frac{\mathbf{A}}{\sqrt{k}}\right). \tag{6}$$

Then, we aggregate all word features with regard to each region on the basis of normalized matrix $\bar{\mathbf{A}}$:

$$\boldsymbol{U} = \bar{\mathbf{A}} \cdot \boldsymbol{W}, \tag{7}$$

where the ith row of $\boldsymbol{U}$ denotes the interactive textual features corresponding to the ith region. Therefore, $\boldsymbol{U}$ can be used to explore the interaction of information flowing between images and sentences by aligning regions and words.

2) Cross-modal Gating Module: The cross-modal alignment module generates the most corresponding word-level information for each region, and fragment messages passing across the two modalities allow fine-grained cross-modal interactions. However, in practice, not all the learned region-word pairs are perfectly aligned. Therefore, we further propose a cross-modal gating module utilizing a soft gate to control feature fusion intensity adaptively, with the aim to eliminate the influence of negative region-word pairs and further enhance the interactions of cross-modality information.

Specifically, for the ith region feature $\boldsymbol{r}_i$ and the associated textual features $\boldsymbol{u}_i$ with regard to the ith region (the ith row of $\boldsymbol{U}$), we evaluate the extent of alignment by calculating the gate value as:

$$\boldsymbol{g}_i = \sigma(\boldsymbol{r}_i * \boldsymbol{u}_i), \tag{8}$$

where $\sigma(\cdot)$ represents the sigmoid function. If a region is well aligned with the corresponding sentence words, a high gate value will be obtained to take full advantage of aligned pairs. Conversely, if a region is not well aligned with the sentence words, a low gate value will be obtained to filter out negative information. We then utilize the gate value to control how much correspondence information should be passed:

$$\boldsymbol{c}_i = \boldsymbol{g}_i * [\boldsymbol{r}_i, \boldsymbol{u}_i], \tag{9}$$

$$\boldsymbol{o}_i = \text{ReLU}(\mathbf{W}_o \boldsymbol{c}_i + \boldsymbol{b}_o), \tag{10}$$

$$\boldsymbol{z}_i = \boldsymbol{o}_i + \boldsymbol{r}_i, \tag{11}$$

where $\mathbf{W}_o$ and $\boldsymbol{b}_o$ are parameters to be learned. $\boldsymbol{c}_i$ is a fused feature that encourages the fusion to a large extent if the region-word pair is well aligned. On the contrary, $\boldsymbol{c}_i$ is suppressed by a low gate value if the region-word pair is not well aligned, and the region feature $\boldsymbol{r}_i$ is combined to preserve the original information. The cross-modal interaction is further enhanced with $\boldsymbol{c}_i$. Since we employ m regions for an image in this work, $\boldsymbol{Z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_m\}, \boldsymbol{Z} \in \mathbb{R}^{m \times d}$ denotes the fused features, which imply the alignment messages between regions and words.

Finally, we aggregate features $\boldsymbol{Z}$ from m regions to represent the whole input image-text pair based on attention mechanism as:

$$\alpha_z = \text{softmax}(Tran_z(\boldsymbol{Z}))^{\text{T}}, \tag{12}$$

$$\boldsymbol{C} = \alpha_z \boldsymbol{Z}, \tag{13}$$

where $Tran_z$ represents a trainable transformation consisting of linear layers and non-linear activation function. The final fused alignment representation $\boldsymbol{C} \in \mathbb{R}^d$ is obtained for an image-text pair.

Positive *Negative*

Fig. 4. An example to illustrate the impact of context in sentiment analysis. The same object in different contexts may evoke different sentiments. *Flowers* in the left convey positive emotion, while *flowers* in the right evoke negative emotion.

B. Context Information Extraction

Context information plays a significant role in sentiment analysis. The same object may evoke different sentiments in different visual contexts. Fig. 4 shows such an example. In the context of a wedding, flowers in the left image convey positive sentiment. On the contrary, flowers in front of a tombstone evoke negative emotions such as sadness and compassion. Apparently, the interactions between image objects and visual

context jointly determine the overall sentiment. Similarly, textual context also affects sentiment prediction significantly. Taking the word "dog" as an example, it is obvious that two sentences "My dog is dead." and "My dog is lovely." would evoke completely different emotions. Therefore, it is necessary to take context information into account in multimodal sentiment analysis.

To extract visual context, the image $\boldsymbol{I}$ is resized into 224×224 and then fed into CNN to obtain the contextual representation. In this work, we employ ResNet18 [49] pre-trained on ImageNet [54] and remove the top fully connected layer to extract 512-dimentional visual context feature $\boldsymbol{V}$:

$$\boldsymbol{V} = ResNet18(\boldsymbol{I}). \tag{14}$$

By the aforementioned cross-modal alignment module, we obtain the word-level features $\boldsymbol{W} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_n\}$ with context information for an input sentence $\boldsymbol{T}$ with n words. Hence, we average these word representations to produce the textual context feature $\boldsymbol{S}$ in a sentence-level:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i. \tag{15}$$

Considering the complementary role of context information in sentiment analysis, we integrate the single-modal contextual features $\boldsymbol{V}$ and $\boldsymbol{S}$ with the region-word alignment $\boldsymbol{C}$ respectively. The aggregated visual and textual representations are generated through multilayer perceptron (MLP) as:

$$\boldsymbol{F}_1 = \mathrm{MLP}(\boldsymbol{V} \oplus \boldsymbol{C}), \tag{16}$$

$$\boldsymbol{F}_2 = \mathrm{MLP}(\boldsymbol{S} \oplus \boldsymbol{C}), \tag{17}$$

$$\boldsymbol{F} = \lambda \boldsymbol{F}_1 + (1-\lambda)\boldsymbol{F}_2, \tag{18}$$

where $\oplus$ represents the concatenation operation, and λ controls the tradeoff between aggregated visual and textual features. The selection of λ is further discussed in Sec. IV-F. As a result, we obtain the final cross-modal interactive feature $\boldsymbol{F}$ that captures both alignment and context information.

C. Multimodal Sentiment Classification

The objective of multimodal sentiment classification is to predict the sentiment label $y \in \{Positive, Neutral, Negative\}$ of an input image-text pair $(\boldsymbol{I}, \boldsymbol{T})$. Therefore, we feed the feature vector $\boldsymbol{F}$ into a softmax layer for predicting the final sentiment:

$$y = \mathrm{softmax}(\mathbf{W}_f \boldsymbol{F} + \boldsymbol{b}_f), \tag{19}$$

where $\mathbf{W}_f$ and $\mathbf{b}_f$ are learnable parameters. Our model is trained by minimizing the cross-entropy loss with the Adam optimization:

$$\mathcal{L} = -\sum_i \widehat{y}_i \log y_i, \tag{20}$$

where $\widehat{y}_i$ denotes the ground truth sentiment label, and y_i is the output of the softmax layer.

TABLE I
STATISTICS OF THE PROCESSED MVSA DATASETS.

Dataset	Positive	Neutral	Negative	Total
MSVA-Single	2683	470	1358	4511
MSVA-Multiple	11318	4408	1298	17024

IV. EXPERIMENTS

A. Datasets

We evaluate the proposed model on two public multimodal sentiment analysis databases, including MVSA-Single and MVSA-Multiple [55]. **MVSA-Single** consists of 5,129 image-text pairs collected from Twitter. Each pair is labeled by an annotator, who assigns one sentiment (positive, neutral and negative) to the image and text respectively. **MVSA-Multiple** contains 19,600 image-text pairs. Each pair is labeled by three annotators, and the sentiment assignment for image and text of each annotator is independent.

For fair comparison, we preprocess the two datasets following the method in [35], where the pairs of inconsistent image label and text label were removed. Specifically, if one label is positive (or negative) and the other is neutral, we take the sentiment polarity of this pair as positive (or negative). Therefore, we get the new MSVA-single and MSVA-multiple datasets for our experiments, as illustrated in Table I.

B. Implementation Details

In our experiments, the datasets are randomly divided into training set, validation set and test set by the split ratio of 8:1:1. The proposed ITIN is optimized by Adam. The learning rate is initialized as 0.001 and decreased by a factor of 10 every 10 epochs with a weight decay of $1e-5$. Considering the number of samples in two datasets is different, we set the batch size of MVSA-Single to 64 and the batch size of MVSA-Multiple to 128. Our framework is implemented by PyTorch [56]. We use the accuracy and F1-score as the evaluation metrics, which are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{21}$$

$$Recall = \frac{TP}{TP + FN}, \tag{22}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{23}$$

where TP represents True Positive, FP represents False Positive and FN denotes False Negative.

C. Baelines

We compare our model with the following baseline methods. **SentiBank & SentiStrength** [13] detects adjective noun pairs as the mid-level representation of an image based on SentiBank and retrieves the sentiment score of tweet text by SentiStrength. **CNN-Multi** [57] applies two individual CNNs to learn visual and textual features, which are then concatenated as the input of another CNN for sentiment prediction.

TABLE II
COMPARISON OF DIFFERENT METHODS ON MVSA DATASETS.

Method	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
SentiBank&Strength [13]	0.5205	0.5008	0.6562	0.5536
CNN-Multi [57]	0.6120	0.5837	0.6639	0.6419
DNN-LR [58]	0.6142	0.6103	0.6786	0.6633
HSAN [6]	-	0.6690	-	0.6776
MultiSentiNet [35]	0.6984	0.6963	0.6886	0.6811
CoMN [7]	0.7051	0.7001	0.6992	0.6983
MVAN [59]	0.7298	0.7298	0.7236	0.7230
ITIN (Ours)	**0.7519**	**0.7497**	**0.7352**	**0.7349**

TABLE III
ABLATION STUDY RESULTS ON MVSA DATASETS.

Method	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
ITIN w/o Align	0.7132	0.7048	0.7117	0.7107
ITIN w/o Gating	0.7309	0.7299	0.7205	0.7197
ITIN w/o Context	0.7331	0.7320	0.7217	0.7209
ITIN only Context	0.7154	0.7161	0.7158	0.7140
ITIN	**0.7519**	**0.7497**	**0.7352**	**0.7349**

DNN-LR [58] employs pre-trained CNNs for image and text respectively, and uses logistic regression to perform sentiment classification. **HSAN** [6] proposes a hierarchical semantic attentional network on the basis of image captions. For an input image-text pair, the text is processed in a hierarchical structure and the image caption is utilized as semantic information. **MultiSentiNet** [35] extracts object and scene from image as visual semantic information and introduces a visual feature guided LSTM model to extract important words by attention. All these features are then aggregated together to obtain final label. **CoMN** [7] proposes a co-memory network modeling the mutual influences between image and text iteratively to improve multimodal sentiment classification. **MVAN** [59] develops a multimodal sentiment analysis model on the basis of the multi-view attention network, which leverages a memory network module to obtain semantic image-text features iteratively. This approach achieves the state-of-the-art performance compared with other image-text multimodal sentiment analysis studies.

D. Comparison with the state-of-the-art

The performance comparisons between the proposed ITIN and the baselines as measured by accuracy and F1-score on the MVSA-Single and MVSA-Multiple datasets are shown in Table II.

From the results in Table II, we have the following observations: (1) SentiBank & SentiStrength delivers the worst performance because it is based on hand-crafted features; (2) deep learning methods, CNN-Multi and DNN-LR, perform better than SentiBank & SentiStrength, which is mainly benefited from the powerful representation ability of deep neural networks; (3) since RNN can model the text context better, HSAN achieves better performance compared with CNN-based approaches; (4) considering the influence of visual information on text, MultiSentiNet achieves better performance; (5) CoMN learns the mutual influences between visual and textual contents iteratively, and is slightly superior to MultiSentiNet; and (6) MVAN not only leverages image features from object and scene viewpoint, but also conducts interactive cross-modal learning, achieving the best performance among all the baselines.

On the MVSA-Single dataset, our model outperforms the existing best model MVAN by a margin of 2.21% and 1.99% in terms of accuracy and F1-score respectively. For the MVSA-Multiple dataset, our model achieves performance improvement of 1.16% and 1.19% respectively. Overall, these results demonstrate the advantage of the proposed ITIN for multimodal sentiment analysis. The performance improvements benefit from the superiority of ITIN. First, with the proposed cross-modal alignment module and cross-modal gating module, the interactions between image and text can be captured thoroughly at a finer level. Second, the joint consideration of the individual-modal context features can further extract useful and complementary information for more accurate sentiment prediction.

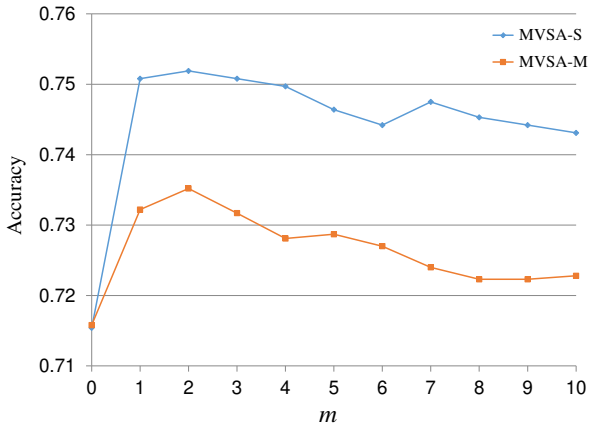
E. Ablation Studies

To further validate the effectiveness of each proposed module, we conduct several ablation experiments on two MVSA datasets in this section. We remove the cross-modal alignment module, the cross-modal gating module and individual-modal contextual features on the basis of ITIN model respectively, which are denoted as "ITIN w/o Align", "ITIN w/o Gating" and "ITIN w/o Context" in Table III. In addition, we also conduct an experiment with individual-modal context representations alone, which is represented as "ITIN only Context". The results of these studies are presented in Table III.

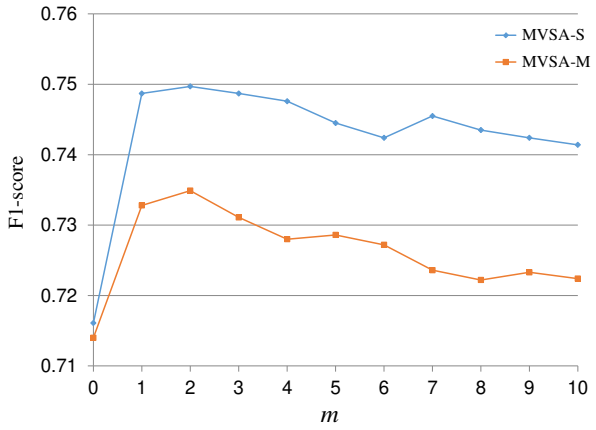
From these results, we can observe that: (1) the proposed ITIN consisting of all modules achieves the best performance on two datasets. The removal of any one module would lead to suboptimal prediction results; (2) ITIN w/o Align is worse than ITIN, which demonstrates the effectiveness of capturing latent alignment between image regions and sentence words; (3) compared with ITIN, the inferior results of ITIN w/o Gating indicate that cross-modal interactions can be further enhanced with the help of adaptive gating mechanism; (4) ITIN w/o Context achieves worse performances than ITIN, verifying the complementary role of context information on sentiment prediction; and (5) from the comparison results of ITIN w/o Context and ITIN only Context, it is obvious that the fine-level cross-modal interactions significantly benefit the multimodal sentiment analysis. From the above observations, we can draw the conclusion that each proposed module is indispensable and jointly makes contribution to final performance.

F. Hyperparameter Analysis

In our experiment, image region number m and λ in Equation (18) are considered as two key hyperparameters

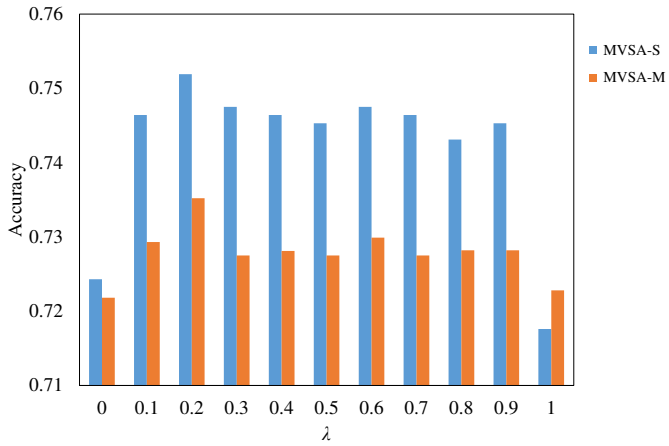


(a) The accuracy performance of ITIN with different m

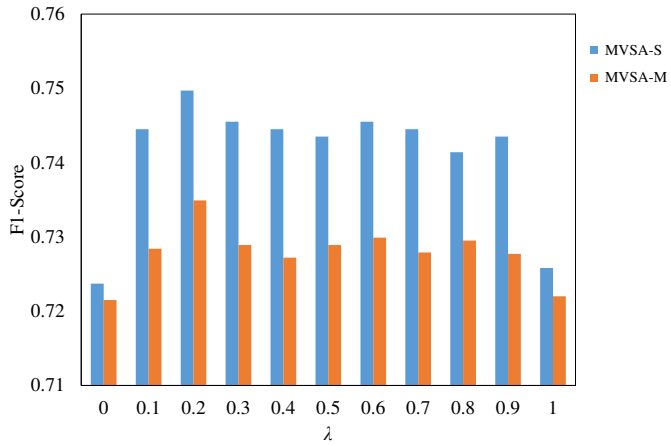


(b) The F1-score performance of ITIN with different m

Fig. 5. Hyperparameter Analysis of different number of image regions m for proposed ITIN.



(a) The accuracy performance of ITIN with different λ



(b) The F1-score performance of ITIN with different λ

Fig. 6. Hyperparameter Analysis of different λ in Equation (18) for proposed ITIN.

influencing the prediction performance. Therefore, we analyze the performance of the proposed method when m and λ are set to various values in this section.

We first analyze the effect of the number of image regions m on the proposed method. Fig. 5 shows the experimental results with different region numbers. MVSA-S denotes results performed on the MVSA-Single dataset and MVSA-M denotes MVSA-Multiple. $m = 2$ can achieve better performance, which is consistent with the aforementioned intuition that human sentiment is evoked mostly by affective regions in an image [8], [9]. Here, $m = 0$ denotes the removal of region-word alignments based on ITIN, and its worst performance demonstrates that cross-modal interactions can be effectively explored with our proposed method. Therefore, from the above discussion, we choose the number of image regions $m = 2$ for all experiments in this paper.

We also conduct experiments when λ in Equation (18) is set to different values. As shown in Fig. 6, setting $\lambda = 0.2$ obtains the best accuracy and F1-score on both MVSA-Single and MVSA-Multiple datasets. We can see that only utilizing visual context features ($\lambda = 1$), or only utilizing textual context representations ($\lambda = 0$) degrades the prediction performance. Therefore, it is necessary to take both visual and textual context information into consideration for the overall sentiment classification. Based on the above analysis, we set $\lambda = 0.2$ in our method.

G. Visual Results

To better understand the rationality of our method, we visualize the aligned region-word pairs of the input image-text pairs learned by our interaction network in Fig. 7. The salient image regions marked with colorful bounding boxes and the corresponding text words in different shades of color are given in the second column. The bigger the weight of the word, the heavier its color. The correspondences between image regions and text words are identified by colors, i.e., red and yellow in the figure. With the proposed cross-modal gating module, as shown in the third column, high gate values are obtained for well-aligned region-word pairs while low gate

Fig. 7. The visualization of the alignment between image regions and sentence words by our ITIN model. The column (a) shows the input image-text pairs. In the column (b), the salient regions are outlined in color, and corresponding text words with different weight are in different shades of colors. The column (c) gives illustration of gate values for matched or mismatched region-word pairs with the proposed cross-modal gating module. Well region-word alignments obtain high gate values and vice versa.

values are generated for mismatched ones. For example, we can observe that the detected region "white dog" with its corresponding word-level text has a gate value of 0.9846 in the second image, while the region "small tree" misaligned with corresponding textual words obtains a gate value of 0.1350 in the first image. By this means, the positive alignment information will be utilized to a large extent and the negative and irrelevant correspondence information will be suppressed during sentiment prediction.

V. Conclusion

In this paper, we have proposed an image-text interaction network (ITIN) for multimodal sentiment analysis. The proposed cross-modal alignment module explores the latent alignment information between image regions and text words. Besides, the negative influences of misaligned region-word pairs can be further filtered out through the proposed cross-modal gating module. With the combined effect of these two modules, cross-modal interactions can be captured at a finer level for overall sentiment classification. In addition, the same object or the same word may evoke different emotions in different context. It is also necessary to take both visual and textual context information as complementary information for achieving more accurate prediction performance. Experimental results and comparisons demonstrate that the proposed ITIN outperforms the state-of-the-art approaches consistently on public databases.

While very encouraging performance has been achieved, one deficiency of our model is that the proposed fine-level interactions would show limited effect when corresponding text dose not describe the affective image regions. In such case, the prediction performance would only rely on individual-modal contextual representations. As future work, we plan to explore mutual influences between image and text at both global and local level simultaneously, which could utilize cross-modal interactions more comprehensively regardless of whether the image-text pairs have corresponding information or not. Furthermore, considering that most of the current multimodal methods focus on sentiment classification, we also plan to investigate multimodal continuous emotion intensity in future work, which could provide richer information on sentiment.

References

[1] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 617–663, 2019.

[2] Y. Gao, Y. Zhen, H. Li, and T.-S. Chua, "Filtering of brand-related microblogs using social-smooth multiview embedding," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2115–2126, 2016.

[3] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[4] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[5] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, 2020, doi:10.1109/TMM.2020.3003648.

[6] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, 2017, pp. 152–154.

[7] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 929–932.

[8] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2020.

[9] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[10] H. Zhu, L. Li, H. Jiang, and A. Tan, "Inferring personality traits from attentive regions of user liked images via weakly supervised dual convolutional network," *Neural Processing Letters*, pp. 2105–2121, 2020.

[11] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 83–92.

[12] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 47–56.

[13] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 223–232.

[14] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013, pp. 1–8.

[15] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 1025–1028.

[16] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[17] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, pp. 381–388.

[18] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network." in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017, pp. 3266–3272.

[19] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 231–237.

[20] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[21] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[22] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[23] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 142–150.

[24] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the International Conference on Computational Linguistics*, 2010, pp. 36–44.

[25] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751.

[26] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015, pp. 1556–1566.

[27] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1422–1432.

[28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2016, pp. 1480–1489.

[29] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 606–615.

[30] L. F. Barrett, K. A. Lindquist, and M. Gendron, "Language as context for the perception of emotion," *Trends in Cognitive Sciences*, vol. 11, no. 8, pp. 327–332, 2007.

[31] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, 2008, pp. 92–103.

[32] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[33] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, "Microblog sentiment analysis based on cross-media bag-of-words model," in *Proceedings of International Conference on Internet Multimedia Computing and Service*, 2014, pp. 76–80.

[34] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2016, pp. 13–22.

[35] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis," in *Proceedings of the ACM on Conference on Information and Knowledge Management*, 2017, pp. 2399–2402.

[36] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M. C. H. Chua, and M. Liu, "An image-text consistency driven multimodal sentiment analysis approach for social media," *Information Processing & Management*, vol. 56, no. 6, p. 102097, 2019.

[37] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 873–883.

[38] Q.-T. Truong and H. W. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 305–312.

[39] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 371–378.

[40] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proceedings of the IEEE International Conference on Data Mining*, 2017, pp. 1033–1038.

[41] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

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

[43] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[44] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2247–2256.

[45] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based Systems*, vol. 161, pp. 124–133, 2018.

[46] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[48] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguisticss*, 2018, pp. 4171–4186.

[51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations*, 2015.

[52] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[53] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5764–5773.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[55] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *Proceedings of the International Conference on Multimedia Modeling*, 2016, pp. 15–27.

[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proceedings of the Conference on Neural Information Processing Systems Worshop*, 2017.

[57] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, 2015, pp. 159–167.

[58] Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 41, pp. 1–11, 2016.

[59] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, 2020, doi:10.1109/TMM.2020.3035277.

Tong Zhu received the B.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China. Her research interests include multimodal sentiment analysis and affective computing.

Leida Li received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search (ROSE) Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the School of Artificial Intelligence, Xidian University. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as an SPC for IJCAI from 2019 to 2021, the Session Chair for ICMR in 2019 and PCM in 2015, and a TPC Member for CVPR in 2021, ICCV in 2021, AAAI from 2019 to 2021, ACM MM from 2019 to 2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 2016. He is also an Associate Editor of the Journal of Visual Communication and Image Representation and the EURASIP Journal on Image and Video Processing.

Jufeng Yang received the Ph.D. degree from Nankai University, Tianjin, China, in 2009. He was a visiting scholar with the Vision and Learning Lab, University of California, Merced, USA, from 2015 to 2016. He is currently an Associate Professor with the College of Computer Science, Nankai University. His research is in the field of computer vision, machine learning, and multimedia.

Sicheng Zhao received his Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2016. He is a postdoctoral research scientist at Columbia University, New York, New York, 10032, USA. He was a visiting scholar at the National University of Singapore from 2013 to 2014, a research fellow at Tsinghua University from 2016 to 2017, and a research fellow at the University of California, Berkeley from 2017 to 2020. His research interests include affective computing, multimedia, and computer vision.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of the IEEE Transactions on HumanC Machine Systems and the IEEE Transactions on Multimedia.

Jiansheng Qian received the B.S. degree from Xidian University, Xian, China, in 1985, the M.S.degree from Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xian, China, in 1988, and the Ph.D. degree in 2003. He is currently a Professor with the China University of Mining and Technology, Xuzhou, China. His research interests include the areas of signal processing for communication, broadband network technology and applications, and coal mine communication and monitoring.