

Image-Text Multimodal Emotion Classification via Multi-View Attentional Network

Xiaocui Yang , Shi Feng, Daling Wang, and Yifei Zhang

Abstract—Compared with single-modal content, multimodal data can express users' feelings and sentiments more vividly and interestingly. Therefore, multimodal sentiment analysis has become a popular research topic. However, most existing methods either learn modal sentiment feature independently, without considering their correlations, or they simply integrate multimodal features. In addition, most publicly available multimodal datasets are labeled by sentiment polarities, while the emotions expressed by users are specific. Based on this observation, in this paper, we build a large-scale image-text emotion dataset (i.e., labeled by different emotions), called TumEmo, with more than 190,000 instances from Tumblr.¹ We further propose a novel multimodal emotion analysis model based on the Multi-view Attentional Network (MVAN), which utilizes a memory network that is continually updated to obtain the deep semantic features of image-text. The model includes three stages: feature mapping, interactive learning, and feature fusion. In the feature mapping stage, we leverage image features from an object viewpoint and a scene viewpoint to capture effective information for multimodal emotion analysis. Then, an interactive learning mechanism is adopted that uses the memory network; this mechanism extracts single-modal emotion features and interactively models the cross-view dependencies between the image and text. In the feature fusion stage, multiple features are deeply fused using a multilayer perceptron and a stacking-pooling module. The experimental results on the MVSA-Single, MVSA-Multiple, and TumEmo datasets show that the proposed MVAN outperforms strong baseline models by large margins.

Index Terms—Memory network, multi-view attention mechanism, social media, multimodal emotion analysis.

I. INTRODUCTION

THE increased use of mobile Internet and smartphones has provided researchers with massive archives of multimodal user-generated content (e.g., text, image, and video) on diverse topics and entities.¹ The tasks of extracting and analyzing the

sentiments embedded in these data have not only attracted substantial attention from academic communities [1], [2] but also have broad commercial application prospects. Although the existing studies have achieved promising results, the literature largely focuses on tasks that use single-modal data, such as text sentiment polarity classification [3] and image emotion recognition [4], but ignore the vivid and complementary sentiment information in multimodal data.

Recently, sentiment analysis has outgrown the traditional single modality approach, and multimodal sentiment tasks that consider emotional features from different modalities simultaneously have emerged as a popular research topic in the multimedia mining community [5], [6]. Early studies required hand-crafted feature engineering for each modality, which is a potentially biased and labor-intensive method [7]. As social media platforms such as Twitter, Tumblr and Weibo aggregate increasingly large amounts of data, the performances of proposed approaches have made significant progress due to the development of deep learning models. Pérez-Rosas *et al.* used a video dialogue dataset, extracted linguistic, acoustic and visual features from video comments, and leveraged SVM classifier to classify emotions expressed in conversation-level visual data streams [8]. Poria *et al.* utilized a multiple-kernel learning classifier to conduct multimodal sentiment analysis on short video segments and proposed an LSTM classification model based on utterances, which enabled contextual information to be obtained from surrounding utterances [9]–[11]. Xu *et al.* proposed a bidirectional multilevel attention (BDMLA) model to leverage complementary information in image and text data to conduct image-text sentiment classification [12].

Although some outstanding deep learning models are available for multimodal sentiment analysis, most existing methods treat the representation learning process of each modality separately and fuse the learned multimodal features at a higher level of the neural network. Furthermore, the cross-modality interactions between different modalities, such as images and text, have received relatively little attention. In this paper, we focus on multimodal emotion analysis for image-text pairs in social media posts. Some motivating examples are illustrated in Table I. In image A, the man's facial expression, the fire, and the word "angry" in the text all imply anger; the text and image content are complementary. In image D, fear is expressed by the girl's facial expression, and the dark environment accentuates this emotion. Similarly, the image of people kissing by the lake expresses love, and the word "romantic" further implies love. In the last example, the teary eyes, facial expressions and rainy

Manuscript received November 20, 2019; revised April 22, 2020 and October 9, 2020; accepted October 24, 2020. Date of publication November 2, 2020; date of current version November 18, 2021. The work was supported in part by the National Key R&D Program of China under Grant 2018YFB1004700 and in part by the National Natural Science Foundation of China under Grant 61872074 and 61772122. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. (Corresponding author: Shi Feng.)



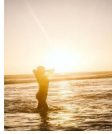




The authors are with the School of Computer Science, and Engineering, Northeastern University, Shenyang 110169, China (e-mail: 2378211148@qq.com; fengshi@cse.neu.edu.cn; wangdaling@cse.neu.edu.cn; zhangyifei@cse.neu.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TMM.2020.3035277>.

Digital Object Identifier 10.1109/TMM.2020.3035277

¹Tumblr is a microblogging and social networking website founded by David Karp in 2007 that is currently owned by Automattic. The service allows users to post multimedia content and other content to short-form blogs. Users can follow other users' blogs.

TABLE I
EXAMPLES OF MULTIMODAL IMAGE-TEXT DATA LABELED BY EMOTIONS

Num	A	B	C	D	E	F	G
Image							
Text	Now you've done it you're fate is sealed #angry #alternativeboys #grungeboy #pissedoff	reaction, reactions, bored, waiting, the Big Lebowski, classic reaction, Big Lebowski	New free stock photo of beach, bright, calm.	Miedito #fineartphotography #blancoynegro #artisticphoto #fear #suspense	Romantic dating. Please click link to join for seeking your rich life partner.	Happy New Year, indeed. 2019 #vacation #travel #thankful #beautiful (at Manjuyod White Sandbar)	New trending GIF tagged sad, rain, emotions, emo, in your feelings via Giphy https://giphy.com/2hNOrOE
Label	Angry	Bored	Calm	Fear	Love	Happy	Sad

environment all express sadness. We can make the following observations based on the above examples.

First, the emotions are not isolated to a single data modality; on the contrary, the emotions in the texts and images are complementary and express the users' sentiments and feelings. However, in the literature, each modal feature is usually modeled separately, and the cross-modal interactions between image and text are ignored [13], [14].

Second, when looking at an image, people usually focus on the part of the image in which they are interested, rather than considering the entire image content equally. Similar to sentiment words in the text, from different views, some of the objects and scenes in the images, such as the girl's smile and the sunshine in F of Table I, are good indicators of emotion that are crucial to the task. Although some methods have successfully utilized object or scene features for sentiment analysis [12], [15], none have considered multi-view features in a unified framework.

Third, the multimodal emotion analysis task (i.e., Table I) is much more difficult than the multimodal sentiment polarity analysis task, not only because there are more emotion categories and the existing models have defects but also because of the lack of large training datasets for multimodal deep learning models.

To tackle these challenges, in this paper, we propose a novel Multi-view Attention Network (MVAN) to achieve robust and accurate multimodal emotion analysis. MVAN consists of three stages, i.e., feature mapping, interactive learning, and feature fusion, and it explores cross-modal interactions and considers the mutual reinforcement between text and image. By observing the image from multiple views or different feature subsets [16], [17], e.g., the image object view and the scene view, we can capture various beneficial features for multimodal emotion analysis. In the feature mapping stage, local object features and scene features are extracted from the image to obtain deep semantic features from a multi-view perspective. In the interactive learning stage, we adopt the image-text interactive learning mechanism. Specifically, the text features are learned from the self-influence of the text under the guidance of image object and scene features. Similarly, the text features help the model learn both the image object features and the scene features. In the feature fusion stage, the four learned features are concatenated and then, to improve the accuracy and F1-score of multimodal emotion analysis, the

features are deeply fused through a multilayer perceptron and a stacking-pooling module. Because no publicly available dataset exists for multimodal emotion analysis, we crawled Tumblr to obtain a large-scale dataset of text-image pairs and used the distant supervision method to label the obtained data. The result is a multimodal emotion analysis dataset named TumEmo, as shown in Table I. The experimental results on the publicly available MVSA-Single, and MVSA-Multiple datasets [18] and on the TumEmo dataset show that the methods proposed in this paper perform satisfactorily on the different multimodal classification tasks.

The main contributions of this paper are as follows:

- We create a image-text dataset (TumEmo) labeled by emotion for multimodal emotion analysis.
- We propose a novel model named MVAN, for multimodal emotion classification and polarity classification.
- Our model outperforms existing methods for multimodal emotion prediction and polarity prediction on different datasets.

The remainder of this paper is organized as follows. In Section II, we survey the related works, including studies on single-modal sentiment analysis, multimodal sentiment analysis, and multimodal fusion. In Section III we present the proposed MVAN model. In Section IV, we report the results of an experimental validation of our approach. Finally, in Section V, we conclude the paper.

II. RELATED WORK

In this section, we briefly review previous studies on text sentiment analysis, image sentiment analysis, multimodal sentiment analysis and multimodal feature fusion.

A. Text Sentiment Analysis

Text sentiment analysis involves extracting and analyzing embedded sentiments or emotions in text and has been widely studied in recent years. Two main types of text sentiment analysis methods exist: lexicon-based approaches and machine learning approaches [19]. The lexicon-based approaches usually employ sentimental words or phrases and rules such as sentiment

inversion and reinforcement to evaluate the polarity of sentences. Taboada *et al.* proposed a method called the Semantic Orientation CALculator (SO-CAL), which uses a dictionary of annotated words with semantic tendencies (both polarity and intensity) and incorporates intensification and negation [20]. Hamouda *et al.* leveraged a sentiment lexicon called SentiWordNet to classify reviews by assigning positive, neutral and negative sentimental scores to each word [21]. In early studies using machine learning approaches, sentiment analysis was considered a text classification problem, and features were designed to improve the results. Pang *et al.* used machine learning methods such as naive Bayes, SVM and maximum entropy models to classify movie reviews [22]. Wang *et al.* proposed a graph model to identify emotions at the hashtag level [23].

Recently, deep learning models have helped to alleviate labor-intensive feature engineering work and improve the final performance. Kim was the first to use a convolutional neural network (CNN) for text sentiment classification [24]. Tang *et al.* employed a CNN and a recurrent neural network (RNN) to improve document-level sentiment classification [25]. Chen *et al.* proposed a novel method that used negative and intensive sentiment supplementary information for sentiment classification [26].

B. Image Sentiment Analysis

Image sentiment analysis is primarily intended to explore the emotions associated with images. In general, the visual features used for image sentiment analysis can be divided into three categories, low-, mid-, and high-level features. For the Flickr dataset, Siersdorfer *et al.* applied the SentiWordNet thesaurus to extract the emotional values of images from metadata such as the URL, image resolution, title, description, and tags [27]. Borth *et al.* proposed SentiBank to detect the presence of 1,200 ANPs (where ANPs are considered to be mid-level image features) in an image [28].

Deep learning models have also achieved promising results for image sentiment analysis tasks. Xu *et al.* utilized a deep convolutional neural network to analyze image sentiment [29]. Song *et al.* proposed a novel visual attention sentiment network that integrated multilevel visual attention into the CNN sentiment classification framework [30]. You *et al.* studied the impact of local image areas on image sentiment analysis [31].

Simonyan *et al.* proposed the family of very deep convolutional networks named VGG-* to improve the classification accuracy of large-scale image recognition (ImageNet) [32]. Zhou *et al.* described the Places Database and provided the Places-CNNs dataset for scene classification purposes [33].

Text sentiment analysis and image sentiment analysis using different approaches have achieved good results. However, while the above methods process only single-modal data, most social media data are multimodal data.

C. Multimodal Sentiment Analysis

Multimodal sentiment analysis makes full use of data from different modes to make accurate sentiment predictions. You *et al.* proposed a cross-modal consistency regression (CCR) model that used a consistent regression model with visual and

text features to train a final sentiment classifier [34]. You *et al.* utilized a tree-based RNN with attention mechanisms for image-text sentiment analysis [35]. Xu proposed a hierarchical semantic attention network model (HSAN) based on image captions. The caption corresponding to an image is used as additional information for multimodal sentiment analysis [13]. Xu *et al.* later proposed MultiSentiNet, which extracts deep semantic features from an image and uses them to guide the learning of text features [14]. Xu *et al.* also proposed a common storage network for multimodal sentiment analysis to iteratively model text and image content and thereby consider the interaction between text and image [15]. Hu *et al.* conducted multimodal sentiment analysis by combining image sentiment with text sentiment using a deep neural network [36]. Huang *et al.* proposed a deep multimodal attention fusion model (DMAF) that used multiple attention mechanisms and mixed fusion methods for image-text sentiment analysis [19]. Ji *et al.* proposed a novel bi-layer multimodal hypergraph learning (Bi-MHG) scheme for multimodal sentiment analysis using text, visual and emoticon modalities [37]. Zadeh *et al.* proposed a tensor fusion network to model intra- and intermodality dynamics [38]. Majumder *et al.* proposed a hierarchical feature fusion strategy that first performs bimodal data fusion (i.e., video-audio (VA), video-text (VT) and text-audio (TA)) and then performs trimodal fusion (VA-VT-TA) [39]. Poria *et al.* proposed a recurrent model that captured the contextual information between utterances and employed an attention mechanism [11]. In other multimodal tasks, adversarial attention networks [40] and linked multimodal data [41] were introduced. Huang *et al.* combined a visual-semantic attention model with an adversarial learning model in an integrated learning framework to learn a joint multimodal representation [40]. Huang *et al.* employed a correlational multimodal variational autoencoder (CMVAE) and a triplet network to learn a unified representation for the multiple modalities of social images [41].

However, the existing studies have some limitations for image-text multimodal sentiment analysis. First, most current methods do not properly consider the deep semantic features of images that can function as potential emotion indicators from different viewpoints, such as the object viewpoint and the scene viewpoint [37]. Second, because images are more abstract and subjective than text, most research emphasizes the text and neglects the relationship between text and image. Deep learning models are heavily dependent on large-scale training data. However, most existing datasets used for multimodal sentiment analysis are labeled by only *positive*, *negative*, and *neutral* labels [13]–[15]. A few small video conversation datasets with emotion labels are available [11], [38], [39]. To bridge this research gap, we construct a large-scale image-text multimodal dataset (TumEmo) of social media data for multimodal emotion analysis.

D. Multimodal Fusion

Effectively integrating image and text features to improve the final classification results plays a crucial role in multimodal sentiment analysis. Poria *et al.* focused mainly on using audio, visual and text information for multimodal sentiment analysis;

the and paper reports on the available multimodal datasets, discusses various approaches for solving the multimodal sentiment recognition problem, and summarizes various multimodal fusion methods [42]. At present, three popular types of modal fusion exist [19], [42]: early, intermediate and late fusion.

Early fusion [8] combines input-level features by calculating the point sum or dot product between the corresponding position elements of vectors or by concatenating the input vectors to prepare features for subsequent use by machine learning algorithms. However, the early fusion approaches may lose the context and time dependence within the modal, and they cannot capture complementary information in intermodal data effectively, resulting in large redundancies.

Late fusion [19] first independently models each modality and then integrates the results from multiple classifiers using a fusion method, such as the majority rule method or the weighted average method. In late fusion, it is assumed that the various modalities are independent; the classifier for each modality is trained independently and thus cannot capture the correlations between modalities.

Intermediate fusion is mainly conducted in the middle of a neural network, such as a compact bilinear pooling network [43] or a tensor fusion network [38]. Intermediate fusion leverages a network sharing layer to merge neural units from multimodal data and the middle layer shares parameters to fuse different modal features.

Most studies model each modality independently and fuse modal features at a high level. However, multimodal features are heterogeneous, and simple fusion cannot sufficiently explore information from different modalities. In this paper, we leverage intermediate fusion to deeply fuse the modal features in the middle layer of the network using a stacking-pooling module.

III. MULTI-VIEW ATTENTIONAL NETWORK FOR MULTIMODAL SENTIMENT ANALYSIS

In this section, we describe the proposed Multi-view Attention Network (MVAN) for multimodal sentiment analysis in detail. We formulate the problem in the first part of this section. Then, we describe the framework of the model, which includes three modules: a feature mapping module, an interactive learning module, and a feature fusion module. Finally, the model algorithm is presented. For convenience, multimodal polarity analysis and emotion analysis are collectively referred to as multimodal sentiment analysis.

A. Problem Formalization

The image-text multimodal sentiment prediction problem is defined as follows. Suppose that T and M represent a text sample space and an image sample space, respectively, where one string of text and its corresponding image constitute an instance. Each instance is associated with a sentiment label L^i . That is, each instance is a triple containing text, an image, and a label and can be expressed as follows:

$$Ins = \{(T^0, M^0, L^0), (T^1, M^1, L^1), \dots, (T^i, M^i, L^i), \dots, (T^{u-1}, M^{u-1}, L^{u-1})\}, \quad (1)$$

where Ins is a set of instance triples, T^i represents the text content, M^i represents the visual content, L^i is the label of the image-text pair in the i^{th} instance, and u represents the number of instances.

The goal of multimodal sentiment prediction is to learn a mapping function $f : (T, M) \rightarrow L$ from the multimodal training dataset $\{(T^i, M^i, L^i) | 0 \leq i \leq u - 1\}$. For polarity classification, $L^i \in \{Positive, Neutral, Negative\}$; for emotion classification, $L^i \in \{Angry, Bored, Calm, Fear, Happy, Love, Sad\}$.

B. Multi-View Attentional Network

To explore the complementarity between the text and image information, we propose a multimodal sentiment analysis model based on a Multi-view Attentional Network (MVAN) to model the interaction between text and image. The framework of MVAN is shown in Fig. 1, and the corresponding algorithm is shown in Algorithm 1. In Fig. 1, the model is divided into three stages: a feature mapping stage, an interactive learning stage, and a feature fusion stage. The details of the multi-view attentional network are described below.

1) *Feature Mapping*: For the text data, a pretrained Glove model [44] is first applied to the Twitter dataset to obtain word embeddings.

$$E^i = T^i W_e, E^i \in R^{l_T \times n}, \quad (2)$$

where T^i is the text content of the i^{th} instance, E^i is the word embedding of the i^{th} instance, W_e represents a weighted matrix, l_T is the length of each string of text, and n contains the dimensions of word vectors. Because both the local text features and the long-term text information are effective for sentiment analysis, we use a CNN to capture local information and a BiLSTM model to capture long-term text information. We use convolution windows with convolution kernels of 2 and 3 to obtain the local features consisting of 2-gram and 3-gram text strings. Then, we adopt max pooling to obtain the most significant local features under different convolution windows.

$$F_{CNN}^i = f_{CNN}(E^i; \theta_t^c). \quad (3)$$

Here, f_{CNN} represents the CNN operation, including the convolution and max pooling operations. θ_t^c is a CNN parameter.

Then, a sentence vector representation of the text is obtained through the BiLSTM:

$$H^i = f_{Bi}(E^i; \theta_t^{Bi}), H^i \in R^{l_T \times d}, \quad (4)$$

where $H^i = h_0^i, h_1^i, \dots, h_j^i, \dots, h_{d-1}^i$ is the output of the BiLSTM, θ_t^{Bi} holds the BiLSTM parameters, and there are $d/2$ hidden units in the BiLSTM. In each time step, we concatenate the hidden layer output of the forward and backward networks. A specific illustration of text feature extraction is shown in Fig. 2. The text features $F_T^i = [t_0^i, t_1^i, \dots, t_j^i, \dots, t_{l_T-1}^i]$ are shown:

$$F_T^i = f_{concat}(F_{CNN}^i, H^i), F_T^i \in R^{l_T \times D_T}, \quad (5)$$

where D_T is the dimension of the connected text features.

It is assumed that objects in an image can reflect embedded emotions. For example, a bouquet of roses represents positivity

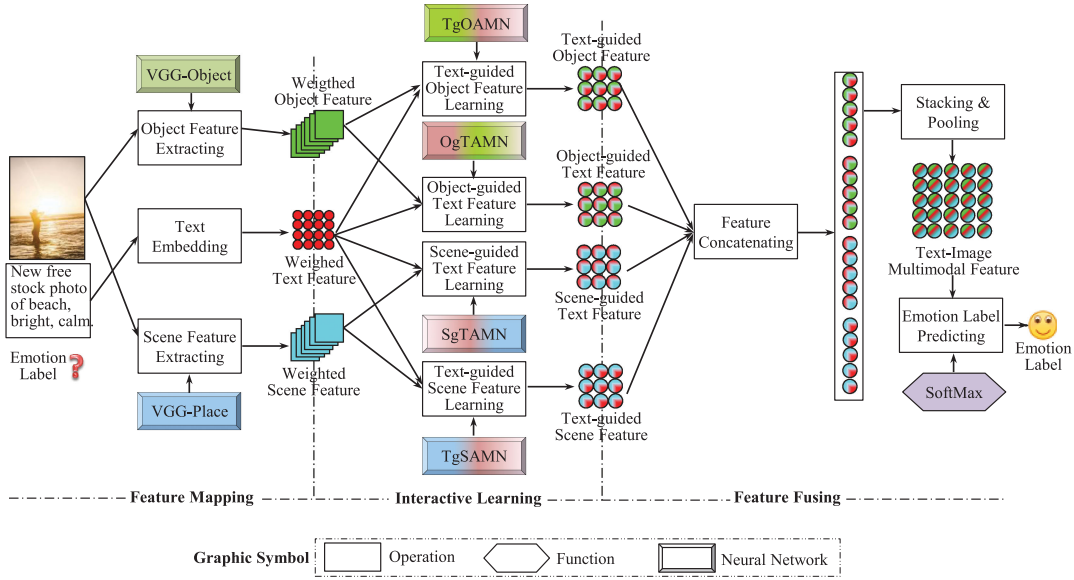


Fig. 1. The framework of the proposed Multi-view Attention Network (MVAN) model for multimodal emotion analysis (predicting emotion labels). We extract image features from both the object view and the scene view. The figure shows the prediction process for an image-text pair without an emotion label. The process employs VGG-Object and VGG-Place models, which are pretrained VGG networks for image processing. TgOAMN, OgTAMN, SgTAMN, and TgSAMN are attentional memory networks trained by our training process used for image object feature learning guided by text features, text feature learning guided by image object features, text feature learning guided by image scene features, and image scene feature learning guided by text features, respectively. The feature fusion module deeply fuses the various features via a stacking-pooling module. The training process includes feature mapping, interactive learning, and feature fusion stages.

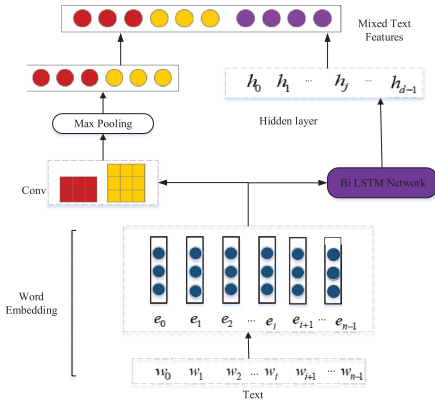


Fig. 2. The text feature extraction process. The upper-left part of the figure shows the process of extracting text local features with a CNN, while the right side shows the process of extracting long-term text features.

and love, while a gun may express negativity and fear. In addition, the scene represented in the image can also reflect emotion. For example, a wedding scene in an image expresses positivity and happiness, while building ruins in a war scene express negativity and fear. Therefore, we extract image features from both the object and scene view to capture sufficient information. In contrast to the global image content, the local features of images are considered to express users' emotions; thus, we use the features from the convolutional layer instead of dense features.

For the image data, we use pretrained VGG-Object [32] and VGG-Place [33] models to extract object-oriented convolutional

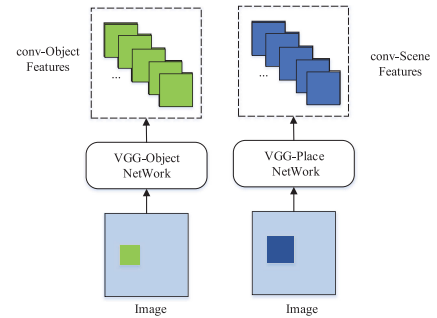


Fig. 3. The process of extracting image convolutional features. For each image, from the object view, we use the VGG-object network to extract the image object features; from the scene view, we use the VGG-place network to extract the image scene features.

features and scene-oriented convolutional features, respectively, as shown in Fig. 3. *Image* represents the set of images, where $Image = \{M^0, M^1, \dots, M^i, \dots, M^{u-1}\}$. The features of each image F_X^i are extracted as follows:

$$F_X^i = f_X(M^i; \theta_X^e), F_X^i \in R^{l_V \times D_V}, \quad (6)$$

where f_X is a visual feature extractor oriented around different views, including the object and scene views, and θ_X^e includes the parameters of different extractors, $X \in \{Object, Place\}$. This extraction process generates l_V conv feature maps, each of which is an $N \times N$ tensor. We flatten each feature map into a $D_V - dimensional$ feature vector. Ultimately, we obtain the object-oriented features $F_O^i = [m_0^i, m_1^i, \dots, m_j^i, \dots, m_{l_V-1}^i]$ and the

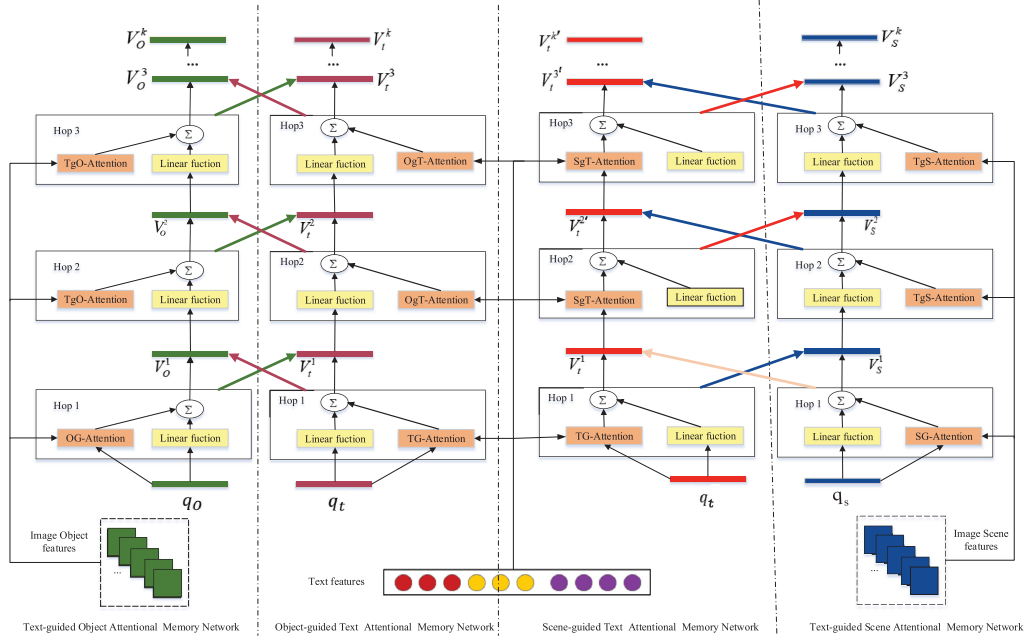


Fig. 4. The interactive learning process using a multi-view attentional network. The network is divided into three parts: a text and image general attention part (including TG-Att, OG-Att and SG-Att); a text-guided image attention memory network part (including TgOAMN and TgSAMN); and an image-guided text attention memory network part (including OgTAMN and SgTAMN).

scene-oriented features $F_S^i = [m_0^i, m_1^i, \dots, m_j^i, \dots, m_{V-1}^i]$ from each image. Here, m_j^i and $m_j^{i'}$ correspond to different parts oriented toward different views.

2) *Interactive Learning*: For a given set of inputs, the memory network considers both the current input and the previous memory to generate a new memory, and each iteration provides relevant information about another modality for the next hop. That is, the interactive stage has the ability to retrieve information that was thought to be irrelevant in previous iterations between different modalities. Each memory network can include multiple hops. Inspired by [15], we utilize the memory network to deepen the interactive learning of different modalities.

Each word in a string of text provides contributes differently to the overall emotional expression. In the same way, when viewing an image, we tend to focus on the part that interests us, rather than the whole image. In other words, the contributions of different pixels in the picture differ in their emotional expression. Therefore, we adopt the attention mechanism to make the model automatically focus on the key parts related to emotional expression to obtain the representations of weighted text features and local features of an image from different views.

The interactive learning module of the memory network mainly explores the auxiliary information between the text and the image to improve the quality of multimodal sentiment analysis. This module iteratively queries the image and text features through the original storage matrix by using the next multihop storage system to explore the relationship between the text and the image. The network is divided into three parts: a general

attention part for the texts and image, a text-guided image attention memory network part and an image-guided text attention memory network part. The details of this module are depicted in Fig. 4.

a) General Attention

For **text general attention**, we stack the high-dimensional text features from the BiLSTM and CNN as an external memory matrix T . Then, T and the query vector q_t are input to the text memory network. First, we obtain a hidden representation of the text through a single-layer perceptron for each string of text:

$$g_j^1 = \tanh(w_{T_{ext}}^1 t_j + b_{T_{ext}}^1). \quad (7)$$

Then, we use the *softmax* function to calculate the attentional weight of the text vector:

$$\alpha_j^1 = \frac{\exp((g_j^1)^T q_t)}{\sum_{i=0}^{l_T-1} \exp((g_i^1)^T q_t)}. \quad (8)$$

Finally, we obtain a weighted text vector representation:

$$v_{T_{ext}}^1 = \sum_{j=0}^{l_T-1} \alpha_j^1 t_j. \quad (9)$$

For **image general attention**, the extracted image features are stacked into an external image-feature memory matrix X . X and an image-object query vector q_x is input to the image memory network. For each image, a hidden representation of the image is obtained through a single-layer perceptron:

$$x_j^1 = \tanh(w_X^1 (m_j \| m_j') + b_X^1). \quad (10)$$

Then, the *softmax* function is used to calculate the attentional weight of the image features:

$$\beta_j^1 \| (\beta_j^1)' = \frac{\exp((x_j^1)^T q_x)}{\sum_{k=0}^{l_V-1} \exp((x_k^1)^T q_x)}. \quad (11)$$

Finally, we obtain the weighted image features:

$$v_X^1 = \sum_{j=0}^{l_V-1} (\beta_j^1 \| (\beta_j^1)') (m_j \| m_j'), \quad (12)$$

where $X \in \{Object, Scene\}$ and $x \in \{Object, Scene\}$ represent different views, and v_X^1 contains the view-oriented weighted image features, including the object-oriented and scene-oriented features when $Hop = 1$. $\beta_j^1 \| (\beta_j^1)' = \beta_j^1$, $m_j \| m_j' = m_j$ when $X \text{ or } x = Object$ and vice versa. Here, $q_t \in R^{D_T}$ and $q_x \in R^{D_V}$ are randomly initialized query vectors learned during training. They help the model consider the important parts related to tasks during training.

b) Text-guided Image Attention Memory Network

Next, we introduce the second part of the deep attention memory network. **The attention memory network learns the image features from different views (TgXAMN)**, including the object view and the scene view—mainly to allow the text to help the model learn the image feature vectors and allow the text features to help the model find the key feature mapping of an image. First, the text vector v_1^{text} is multiplied for each image feature mapping from the views m_i and m_i' ; then, a single-layer perceptron is used to compute the object-oriented hidden representation guided by text with the new visual memory $Memory_{TgX} = Mul(m_j \| m_j', v_1^{text})$:

$$x_j^2 = \tanh(w_X^2 Memory_{TgX} + b_X^2). \quad (13)$$

We adopt the Hadamard product [45], [46], because it retains sufficient interaction information between the text and the image. Second, the *softmax* function is used to calculate attentional weights for the image features guided by the text:

$$\beta_j^2 \| (\beta_j^2)' = \frac{\exp(x_j^2)}{\sum_{k=0}^{l_V-1} \exp(x_k^2)}. \quad (14)$$

Finally, we obtain a new weighted image feature vector:

$$v_X^2 = \sum_{j=0}^{l_V-1} (\beta_j^2 \| (\beta_j^2)') (m_j \| m_j'), \quad (15)$$

where $X \in \{Object, Scene\}$, $x \in \{Object, Scene\}$, and TgX represent that text assists in learning image features from different views. $m_j \| m_j' = m_j$ and $\beta_j^2 \| (\beta_j^2)' = \beta_j^2$ when $X \text{ or } x = Object$, and vice versa.

c) Image-guided Text Attention Memory Network

The attention memory network used to learn text features guided separately by image features from the object view and scene view (XgTAMN) mainly lets the image feature vector help the text feature vector to be learned and to find the important features of text related to image features from different views. The first step is to multiply the image feature vector v_X^1 by each text feature t_j . Then, we use a single-layer perceptron to calculate the text hidden representation guided by the features

from the different image views with the new textual memory $Memory_{XgT} = Mul(t_j, v_X^1)$:

$$g_j^2 \| (g_j^2)' = \tanh(w_{text}^2 Memory_{XgT} + b_{text}^2). \quad (16)$$

Next, the softmax function is used to calculate the attentional weight of text features under the guidance of the image features from the different views:

$$\alpha_j^2 \| (\alpha_j^2)' = \frac{\exp(g_j^2 \| (g_j^2)'))}{\sum_{i=0}^{l_T-1} \exp(g_i^2 \| (g_i^2)'))}. \quad (17)$$

Finally, we obtain a new weighted text feature vector assisted by image features from different views:

$$v_{Text}^2 \| (v_{Text}^2)' = \sum_{j=0}^{l_T-1} (\alpha_j^2 \| (\alpha_j^2)') t_j, \quad (18)$$

where $X \in \{Object, Scene\}$, $x \in \{Object, Scene\}$ and XgT represent the image features from different views that assist in text learning. Here, $g_j^2 \| (g_j^2)' = g_j^2$, $\alpha_j^2 \| (\alpha_j^2)' = \alpha_j^2$, and $v_{Text}^2 \| (v_{Text}^2)' = v_{Text}^2$, when $X \text{ or } x = Object$, and vice versa.

Therefore, to learn the image feature vector guided by the text features of the memory network at the $k - th$ layer, we combine the text representation v_{Text}^{k-1} guided by the object-oriented and scene-oriented image features from the previous hop with the image feature m_j or m_j' to obtain v_X^k by

$$v_X^k = TgXAMN((m_j \| m_j'), (v_{Text}^{k-1} \| (v_{Text}^{k-1})')). \quad (19)$$

Similarly, to learn the text feature vector guided by the image features of the memory network at the $k - th$ layer, we combine the image representations v_X^{k-1} obtained from the object view and the scene view from the previous hop with the text feature t_j to obtain v_{Text}^k or $(v_{Text}^k)'$ as follows:

$$v_{Text}^k \| (v_{Text}^k)' = XgTAMN(t_j, v_X^{k-1}), \quad (20)$$

where $X \in \{Object, Scene\}$, $k \in [2, K]$, where K is the number of memory hops. When $k = 2$, $v_{Text}^{k-1} = (v_{Text}^{k-1})'$.

For the i^{th} instance, the second and third parts of the deep attention memory network when $k > 1$ can be expressed as follows:

$$(V_O^k)^i = TgOAMN(F_O^i, (V_T^{k-1})^i; \theta_O^{TgO}), (V_O^k)^i \in R^{D_V}, \quad (21)$$

$$(V_S^k)^i = TgSAMN(F_S^i, ((V_T^{k-1})')^i; \theta_S^{TgS}), (V_S^k)^i \in R^{D_V}, \quad (22)$$

$$(V_T^k)^i = OgTAMN(F_T^i, (V_O^{k-1})^i; \theta_T^{OgT}), (V_T^k)^i \in R^{D_T}, \quad (23)$$

$$((V_T^k)')^i = SgTAMN(F_T^i, (V_S^{k-1})^i; \theta_T^{SgT}), ((V_T^k)')^i \in R^{D_T}. \quad (24)$$

3) *Feature Fusion*: Effectively integrating text features with image features is an important problem in multimodal sentiment analysis. Inspired by [47], we use the intermediate fusion method to fuse multimodal features through the stacking-pooling module. The specific process is depicted in Fig. 5.

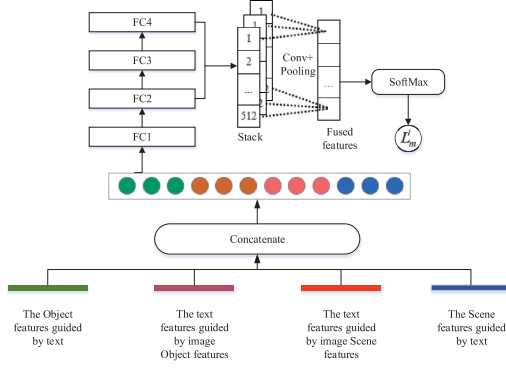


Fig. 5. The process of fusing different features using the stacking-pooling module.

We concatenate the text features (V_T^k and $(V_T^k)'$) with the image features (V_O^k and V_S^k) from the deep attentional memory network to obtain a fused feature V_m^k as follows:

$$(V_m^k)^i = f_{concat}((V_T^k)^i, ((V_T^k)')^i, (V_O^k)^i, (V_S^k)^i). \quad (25)$$

Finally, V_m^k is sent to the multilayer (four-layer) perceptron. To better learn the interactions between modal features, we stack multiple fully connected layers, including FC_2^m , FC_3^m and FC_4^m , and then complete the deep fusion with convolution-pooling operations. Therefore, the stacked layer is a 512×3 matrix, which is convoluted by multiple $1 \times 1 \times 3$ convolution cores and then pooled to obtain the deeply fused features, in which the network weights are shared on the three stacked layers. The convolution and max pooling operations can be expressed as follows:

$$F_m^i = f_{pooling}(f_{conv}((FC_2^m)^i, (FC_3^m)^i, (FC_4^m)^i, \theta_m^{conv})). \quad (26)$$

Finally, the fused features are fed into a *softmax* classifier to obtain the final label, L_m^i :

$$L_m^i = f_{softmax}(F_m^i, \theta_m^{softmax}), L_m^i \in R^C, \quad (27)$$

where θ_m^{conv} and $\theta_m^{softmax}$ are the parameters of the convolution and softmax layers. We adopt the cross-entropy loss function in this study; thus, the model loss function can be expressed as follows:

$$\begin{aligned} J_m(\theta_m^{conv}, \theta_m^{softmax}) &= \sum_{i=0}^{u-1} -\log(L_m^i) \\ &= f_{softmax}(f_{pooling}(f_{conv}(f_{multi-per}([(V_T^k)^i | ((V_T^k)')^i | \\ &\quad (V_O^k)^i | (V_S^k)^i]; \theta_m^{conv})); \theta_m^{softmax}), \end{aligned} \quad (28)$$

where $f_{multi-per}$ is a multilayer perceptron operation, “|” represents the concatenation operation, and the computations of $(V_T^k)^i$, $((V_T^k)')^i$, $(V_O^k)^i$, $(V_S^k)^i$ are shown in Eqs. (21)–(24).

C. The Multi-View Attentional Network Algorithm

Algorithm 1 is the operational flow of the multi-view attentional network.

Algorithm 1: Multimodal Sentiment Analysis Algorithm Based on a Multi-view Attentional Network

Input: Multimodal text-image pairs test dataset:

$\{T, V | \text{length}(T) = u\}$.

Output: The sentiment label set L .

1. Extract the text feature F_T , the object-oriented feature F_O and the scene-oriented feature F_P from the image.
2. $V_T^1 \leftarrow F_T$ by G-Att, $V_O^1 \leftarrow F_O$ by G-Att, $V_S^1 \leftarrow F_P$ by G-Att, when $k = 1$.
3. $V_O^k \leftarrow (F_O, V_T^{k-1})$ by TgOAMN, $V_S^k \leftarrow (F_P, (V_T^{k-1})')$ by TgSAMN, $V_T^k \leftarrow (F_T, V_O^{k-1})$ by OgTAMN, $(V_T^k)' \leftarrow (F_T, V_S^{k-1})$ by SgTAMN, when $k > 1$ and $k = 2$, $V_T^{k-1} = (V_T^{k-1})'$.
4. $F_m \leftarrow (V_T^k, (V_T^k)', V_S^k, V_O^k)$ by the deep fusion module.
5. Update $\theta_m^{softmax}$, θ_m^{conv} using Eq. (28).
6. Repeat steps 2–5 until the accuracy of the validation dataset no longer increases over ten epochs or reaches n times ($n = 100$).
7. Retrieve the predicted label set L_m .

TABLE II
THE NUMBER OF INSTANCES IN THE MVSA DATASET FOR EACH SENTIMENT

Dataset	Positive	Neutral	Negative	All
MVSA-S	2683	470	1358	4511
MVSA-M	11318	4408	1298	17024

IV. EXPERIMENTS

This section describes the experiments conducted on multimodal sentiment analysis of social media content. This section is divided into three parts: experimental data preparation and processing, experimental setup and experimental results and analysis.

A. Experimental Data and Preprocessing

We conduct the multimodal sentiment analysis task using image-text data from social media. We adopt three real-world datasets for these experiments: MVSA-Single, MVSA-Multiple [18], and TumEmo. The statistics for the two MVSA datasets are listed in Table II, showing that the various categories are highly unbalanced and that the MVSA-Single and MVSA-Multiple have different data distributions. If we do not perform sampling, the smaller categories of the datasets will be difficult to study during the training process, and all the data will be classified into the same class (i.e., the classifier will fail). The MVSA datasets are small; therefore, random upsampling method conducted for the smallest category in each MVSA dataset to reduce the impact of data imbalance on the experiment.

While a number of publicly available image-text datasets currently exist, no publicly available dataset containing emotion

TABLE III
THE NUMBER OF INSTANCES OF EACH EMOTION IN THE TUMEMO DATASET

Emotion	Before	After	Percent
Angry	14,544	14,544	100%
Bored	32,283	32,283	100%
Calm	18,109	18,109	100%
Fear	20,264	20,264	100%
Happy	194,107	50,267	25.90%
Love	133,519	34,511	25.85%
Sad	64,173	25,277	39.39%
All	534,931	195,265	36.50%

annotations has been reported. Therefore, to analyze the emotions conveyed by image-text pairs, we crawled Tumblr,² named TumEmo³ containing a large amount of image-text data. The manual annotation of a large dataset is time consuming and laborious; thus, we used a distant supervision method to label the multimodal data. We observed that the data published on Tumblr are usually tagged by users to summarize their personal feelings and emotions. Therefore, we adopted the tags on a post as weak emotion labels of the image-text pair. As shown in Table III, the original TumEmo dataset is very large, and its class distribution is severely imbalanced, which causes a problem during training similar to that discussed for the MVSA datasets. Thus, we also use random downsampling to address the imbalance in TumEmo.

Since text data usually contain many characters that are useless for sentiment analysis, we preprocess the text in the three datasets as follows:

- replace “URL” with the null character;
- replace “@username” with the null character;
- replace punctuation marks that are not useful for sentiment analysis, including brackets, periods, and commas, with the null character;
- replace the content of “#hashtag” with the null character. This deletes any emotional words represented by the hashtag, which can affect the sentiment analysis result.

B. Experimental Setup

In this study, the dataset is divided into a training set, validation set and test set at a ratio of 8:1:1. We utilize Adam as the optimizer method with a learning rate of 0.001. The batch size of MVSA-Single is 32, the batch size of MVSA-multiple is 128, and the batch size of TumEmo is 512. For the four stacked fully connected layers, the number of hidden layer units are 1024, 512, 512, and 512. Then, a two-dimensional operation with a convolution kernel of 1 is carried out, and pooling is conducted. The metrics used in our experiment are accuracy and F1-score.

C. Baselines

We compare our model with the following baseline models. First, to highlight the advantages of multimodal feature fusion,

we evaluate the single-modal sentiment analysis methods, such as CNN-T and OSDA-V in groups I and II of Table IV, as baseline models. Second, we propose two models variants (i.e., SF-M and DSN-M in group III) to validate the superiority of our MVAN-M model. Third, we compare MVAN-M with several strong baseline methods reported in the literature, such as HSAN-M in group IV of Table IV. The notation “(pub)” in the table represents the published results for the MVSA dataset, while the notation “(pro)” in the table represents the results reproduced for the MVSA and TumEmo datasets. In the following list, the suffix “*-T” indicates that a model uses only text features; the suffix “*-V” indicates the use of image features; and the suffix “*-M” indicates multimodal sentiment analysis using both features.

CNN-T [24]: A text sentiment analysis model based on a CNN.

BiLSTM-T [48]: A text sentiment analysis model based on a BiLSTM.

BiACNN-T [49]: A text sentiment analysis model based on a CNN and a BiLSTM with an attention mechanism.

Incep-V [50]: An image sentiment analysis model based on Inception V3.

OSDA-V: An image sentiment analysis model based on double attention to object [32] and scene features [33]. We believe that the image features from different views helps reflect users’ sentiments.

SF-M: A multimodal sentiment model based on simple fusion. This model concatenates the text features from a BiLSTM [48] with the visual features from an InceptionV3 model [50].

DSN-M: This model is an improved version of the Multi-SentiNet method [14] proposed in this paper that leverages the attention mechanism to extract image object features and scene features.

HSAN-M [13]: Xu *et al.* proposed a hierarchical semantic attentional network based on image captions for multimodal sentiment analysis. The network uses a hierarchical structure for the text and uses image captions as visual features.

MultiSentiNet-M [14]: Xu *et al.* proposed a visual-feature-guided LSTM with attention to extract words that were important to text sentiment and then aggregated the text representation, image object features and scene features.

Co-Memory-M [15]: The authors proposed the co-memory network to iteratively model the interactions between visual content and text for multimodal sentiment analysis. This model achieves the current state-of-the-art performance on text-image multimodal sentiment classification tasks.

D. Experimental Results and Analysis

We conduct the experiments with our proposed models and the baselines on three datasets. Note that we conduct multimodal polarity classification on the MVSA-Single and MVSA-Multiple datasets and multimodal emotion classification on TumEmo.

1) *Results of the Baseline Methods and Our Model*: The experimental results of the baseline methods and our model are shown in Table IV, where MVAN-M denotes our model, which

²[Online]. Available: <http://tumblr.com>

³We released the source code and the TumEmo dataset at [Online]. Available: <https://github.com/YangXiaocui1215/MVAN>

TABLE IV
THE METRICS OF ACCURACY AND F1-SCORE ON THREE DATASETS

Model	MVSA-Single		MVSA-Multiple		TumEmo	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
CNN-T	0.6819	0.5590	0.6564	0.5766	0.6154	0.4774
BiLSTM-T	0.7012	0.6506	0.6790	0.6790	0.6188	0.5126
BiACNN-T	0.7036	0.6916	0.6847	0.6319	0.6212	0.5016
Incep-V	0.6362	0.6340	0.6341	0.6207	0.3628	0.1858
OSDA-V	0.6675	0.6651	0.6662	0.6623	0.4770	0.3438
SF-M	0.6892	0.6145	0.6730	0.5499	0.6340	0.5409
DSN-M	0.6940	0.6892	0.6939	0.6926	0.6533	0.6085
HSAN-M (pub)	—	0.6690	—	0.6776	—	—
MultiSentiNet-M (pub)	0.6984	0.6963	0.6886	0.6811	—	—
Co-Memory-M (pub)	0.7051	0.7001	0.6992	0.6983	—	—
HSAN-M (pro)	0.6988	0.6458	0.6796	0.5647	0.6309	0.5398
MultiSentiNet-M (pro)	0.6819	0.6771	0.6815	0.6639	0.6418	0.5962
Co-Memory-M (pro)	0.7108	0.7012	0.7037	0.6961	0.6426	0.5909
MVAN-M (our)	0.7298	0.7298	0.7236	0.7230	0.6646	0.6339

is based on the multi-view attentional network. We can make the following observations. First, our model (MVAN) outperforms the other models in terms of accuracy and F1-score. Second, the multimodal sentiment analysis models perform better than do most of the single-modal sentiment analysis models on all three datasets. Clearly, considering multimodal content is helpful for the sentiment analysis task. However, while some multimodal methods perform slightly worse than do several of the text sentiment analysis models on MVSA-Single and MVSA-Multiple, this phenomenon does not occur on the TumEmo dataset, mainly because it is both large and diverse. A multimodal sentiment analysis model is better able learn the diversity of the different feature modalities and perform more robustly on the TumEmo dataset. The MVSA-Single and MVSA-Multiple contain few samples, while the parameters of the multimodal model are large and difficult to learn, and the image data may hinder learning from text data. Third, the results demonstrate the effectiveness of DSN-M, which is an improved version of MultiSentiment-M. Fourth, we list not only the reported experimental results of the previous models on the MVSA-Single and MVSA-Multiple datasets but also their reproduced results, mainly because we also needed to conduct experiments on the TumEmo dataset. To ensure the fairness of the experiments, we conducted them using our own reproduced code; thus, the results are slightly different from those of the published models.

We find that MVAN-M is superior to the other models because of the interactive learning module of the attention memory network and the deep fusion module. The functions of these two modules are described in detail in previous sections. The attention memory network module mainly improves the interactions of two kinds of modal information; specifically, it allows the text information to help the image features be learned (and vice versa) to achieve the complementary learning of both types of modal information, which improves the results of multimodal sentiment analysis. DSN-M and MultiSentiNet-M consider only the image information to help the text features to be learned, and they ignore the auxiliary effect of text on learning the image features. SF-M and Co-Memory-M use only the image

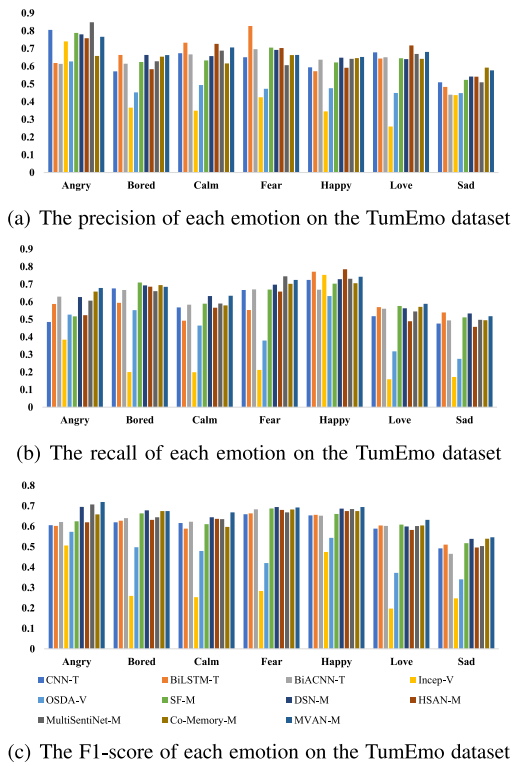


Fig. 6. The experimental results of each category of the TumEmo dataset.

object features; they ignore image scenes, which can also reflect user sentiments. In contrast to our model, all the other tested multimodal methods simply fuse the features of different modalities through concatenation. Instead, we employ the stacking-pooling module to deeply fuse the multimodal content and thus improve the results of multimodal sentiment analysis.

To further demonstrate the advantages of our model, we show its experimental results on each category of the TumEmo dataset in Fig. 6. Because the TumEmo dataset is large and diverse, the models are not severely affected by different factors, and the results are robust. Therefore, we use the experimental results on

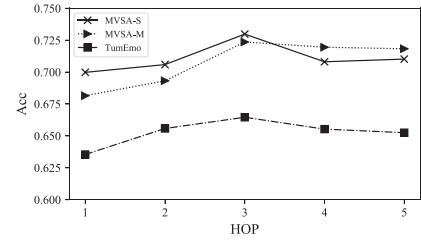
TABLE V
ABLATION EXPERIMENT RESULTS

Datasets	Model	Accuracy	F1
MVSA-Single	MVAN w/o AMN	0.6998	0.6998
	MVAN w/o DF	0.7146	0.6972
	MVAN w/o Scene	0.6843	0.6675
	MVAN w/o Object	0.6820	0.6723
	MVAN-M	0.7298	0.7298
MVSA-Multiple	MVAN w/o AMN	0.6815	0.6786
	MVAN w/o DF	0.7107	0.6651
	MVAN w/o Scene	0.7008	0.7002
	MVAN w/o Object	0.6932	0.6844
	MVAN-M	0.7236	0.7230
TumEmo	MVAN w/o AMN	0.6354	0.6040
	MVAN w/o DF	0.6489	0.5624
	MVAN w/o Scene	0.6560	0.6136
	MVAN w/o Object	0.6527	0.6122
	MVAN-M	0.6646	0.6339

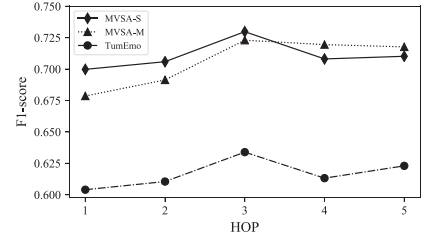
the TumEmo dataset as representative of MVAN's performance. This figure shows that compared with other models, our model (MVAN-M) is competitive on each emotion in terms of precision, recall and F1-score. For the different metrics, our model either outperforms or is competitive with the baseline methods on different emotions. As shown in Fig. 6(a) and Fig. 6(b), other models sometimes outperform our model, but our model performs best in terms of F1-score. Thus, we believe that our model is competitive. Our model has the strongest ability to identify happiness, fear and anger, probably because users express these emotions explicitly via text and images. For sadness, all the models perform worse than they do for the other emotions, possibly because users express sadness in implicit ways.

2) *Ablation Experiments on MVAN*: We conduct ablation experiments on the MVAN-M model to verify the effectiveness of different modules. First, we remove the attention memory module (MVAN w/o AMN) and the deep fusion module (MVAN w/o DF); then, we removed the image scene feature (i.e., we use only the image object features) (MVAN w/o Scene); and finally, we removed the object features of the image (i.e., we use only the image scene) (MVAN w/o Object). The results of these ablation experiments are listed in Table V. Table V shows that the full version of MVAN-M achieves the best performance. The removal of either the interactive learning module or the deep fusion module adversely affects the model results, which indicates that these two modules are effective for multimodal sentiment analysis. This result is achieved mainly because the attention memory network causes both the text and image data to participate in auxiliary learning, which promotes learning during sentiment analysis. The deep fusion module enables the learned text and image features to be effectively fused in a high-dimensional space.

The results of the MVAN w/o DF model on the three datasets are better than those of the MVAN w/o AMN model, indicating that interactive learning with the memory network contributes to the final performance. The ablation experiments removing image features for the three datasets show that the best performance is



(a) The Accuracy of MVAN on different HOPs



(b) The F1-score of MVAN on different HOPs

Fig. 7. Comparison of experimental results for HOP.

achieved when both the image object features and the image scene features of the image are included. However, when one of those views is removed, the performance of the model degrades, which indicates that both the object view and scene view of the image are effective for multimodal emotion analysis. The result is slightly better when the scene view of the image is removed and only the object view (MVAN w/o Scene) is used than when the object view is removed and only the scene view (MVAN w/o Object) is used. We can infer that the view of the image object contributes somewhat more than that of the scene feature to the sentiment analysis task.

3) *The HOP Hyperparameter*: We also conduct experiments under different settings of the hyperparameter HOP. The results in Fig. 7 show that when HOP=3, the accuracy and F1-score reach their maximum values on MSVA-Single, MSVA-Multiple and TumEmo. Therefore, for all the previous experiments reported in this paper, the HOP parameter is set to 3. HOP = 1 denotes that the memory network module of the MVAN is removed in the ablation experiment (MVAN w/o AMN).

V. CONCLUSION

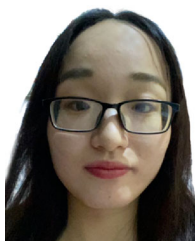
In this study, to perform multimodal emotion analysis of social media content, we built a large image-text dataset (TumEmo) by crawling Tumblr. Based on different views of an image (i.e., the object view and the scene view), we can effectively capture information for the multimodal emotion analysis task. We proposed a novel multimodal emotion analysis model based on a multi-view attention network, which interactively learns text and image features through an attention memory network module. Then, the multimodal feature fusion module is constructed by using a multilayer perceptron and a stacking-pooling module. Both modules significantly improve the quality of multimodal sentiment analysis. The experimental results on two publicly available datasets

and our built dataset demonstrated that our proposed model outperforms strongly competitive baseline models by a large margin. In future work, for interactive learning, we plan to also consider scene-guided objects and object-guided scenes, which could be helpful to multimodal emotion analysis because the meanings of objects can vary depending on the scenes in which they are located. Furthermore, we plan to introduce adversarial learning into multimodal feature fusion.

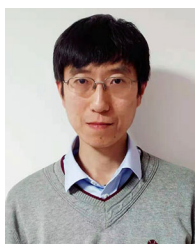
REFERENCES

- [1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018, Art. no. e1253.
- [2] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, no. 2, pp. 617–663, 2019.
- [3] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on multi-feature fusion," *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1245–1259, 2019.
- [4] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.
- [5] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," *Int. J. Serv. Sci., Manag., Eng., Technol.*, vol. 10, no. 2, pp. 38–58, 2019.
- [6] M. Soleymani *et al.*, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.
- [7] F. Wu, Y. Huang, Y. Song, and S. Liu, "Towards building a high-quality microblog-specific chinese sentiment lexicon," *Decis. Support Syst.*, vol. 87, pp. 39–49, 2016.
- [8] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meet. Assoc. Comput. Linguistics*, 2013, pp. 973–982.
- [9] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.
- [10] S. Poria *et al.*, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meet. Assoc. Comput. Linguistics*, 2017, pp. 873–883.
- [11] S. Poria *et al.*, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining*, New Orleans, LA, 2017, pp. 1033–1038.
- [12] J. Xu *et al.*, "Visual-textual sentiment classification with bi-directional multi-level attention networks," *Knowl.-Based Syst.*, vol. 178, pp. 61–73, 2019.
- [13] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, 2017, pp. 152–154.
- [14] N. Xu and W. Mao, "MultiSentiNet: A deep semantic network for multimodal sentiment analysis," in *Proc. ACM Conf. Inf. Knowl. Manag.*, 2017, pp. 2399–2402.
- [15] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 929–932.
- [16] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, vol. abs/1304.5634, 2013.
- [17] S. Sun, L. Mao, Z. Dong, and L. Wu, *Multiview Machine Learning*. Berlin, Germany: Springer, 2019.
- [18] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *Proc. Int. Conf. Multimedia Model.*, 2016, pp. 15–27.
- [19] 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, 2019.
- [20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.
- [21] A. Hamouda and M. Rohaim, "Reviews classification using sentiwordnet lexicon," in *Proc. World Congr. Comput. Sci. Inf. Technol.*, 2011, vol. 23, pp. 104–105.
- [22] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods Natural Lang. Process.-Volume 10*, 2002, pp. 79–86.
- [23] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, 2011, pp. 1031–1040.
- [24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
- [25] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1422–1432.
- [26] X. Chen *et al.*, "Sentiment classification using negative and intensive sentiment supplement information," *Data Sci. Eng.*, vol. 4, no. 2, pp. 109–118, 2019.
- [27] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 715–718.
- [28] 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*, 2013, pp. 223–232.
- [29] C. Xu, S. Cetintas, K. Lee, and L. Li, "Visual sentiment prediction with deep convolutional neural networks," *CoRR*, vol. abs/1411.5731, 2014.
- [30] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.
- [31] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 231–237.
- [32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
- [33] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2017.
- [34] 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*, 2016, pp. 13–22.
- [35] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1008–1017.
- [36] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 350–358.
- [37] 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. 2018.
- [38] 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.
- [39] 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, 2018.
- [40] F. Huang, X. Zhang, and Z. Li, "Learning joint multimodal representation with adversarial attention networks," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1874–1882.
- [41] F. Huang, X. Zhang, J. Xu, Z. Zhao, and Z. Li, "Multimodal learning of social image representation by exploiting social relations," *IEEE Trans. Syst., Man, Cybern.*, pp. 1–13, 2019.
- [42] 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, 2017.
- [43] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 317–326.
- [44] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
- [45] B. Duke and G. W. Taylor, "Generalized hadamard-product fusion operators for visual question answering," in *Proc. 15th Conf. Comput. Robot Vis.*, 2018, pp. 39–46.
- [46] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int. Conf. Learn. Representations (ICLR) (Poster)*, 2016.
- [47] Y. Liu, L. Liu, Y. Guo, and M. S. Lew, "Learning visual and textual representations for multimodal matching and classification," *Pattern Recognit.*, vol. 84, pp. 51–67, 2018.

- [48] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (volume 2: Short papers)*, 2016, pp. 207–212.
- [49] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2267–2273.
- [50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.



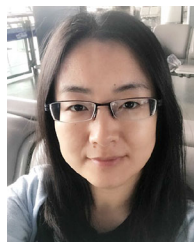
Xiaocui Yang received the M.S. degree from the School of Computer Science and Engineering, Northeastern University, Shenyang, China, in June 2019. Her research interests include multimodal sentiment analysis, machine learning, and text sentiment analysis.



Shi Feng received the Ph.D. degree in computer software and theory from Northeastern University, China. He is currently an Associate Professor with the School of Computer Science and Engineering, Northeastern University, Shenyang, China. He has authored or coauthored more than 20 papers in top-tier journals and conferences, including International Joint Conferences on Artificial Intelligence (IJCAI), The Association for Computational Linguistics (ACL), Special Interest Group on Information Retrieval (SIGIR), and Conference on Empirical Methods in Natural Language Processing (EMNLP). His research interests include sentiment analysis and dialogue systems.



Daling Wang is a Professor with the School of Computer Science and Engineering, Northeastern University, Shenyang, China. Her research interests include social media processing, sentiment analysis, data mining, and information retrieval.



Yifei Zhang is an Assistant Professor with the School of Computer Science and Engineering, Northeastern University, Shenyang, China. Her research interests include image processing and machine learning.