Multi-Level Attention Map Network for Multimodal Sentiment Analysis

Xiaojun Xue, Chunxia Zhang, Zhendong Niu, and Xindong Wu, *Fellow, IEEE*

Abstract—Multimodal sentiment analysis (MSA) is a very challenging task due to its complex and complementary interactions between multiple modalities, which can be widely applied into areas of product marketing, public opinion monitoring, and so on. However, previous works directly utilized the features extracted from multimodal data, in which the noise reduction within and among multiple modalities has been largely ignored before multimodal fusion. This paper proposes a multi-level attention map network (MAMN) to filter noise before multimodal fusion and capture the consistent and heterogeneous correlations among multi-granularity features for multimodal sentiment analysis. Architecturally, MAMN is comprised of three modules: multi-granularity feature extraction module, multi-level attention map generation module, and attention map fusion module. The first module is designed to sufficiently extract multi-granularity features from multimodal data. The second module is constructed to filter noise and enhance the representation ability for multi-granularity features before multimodal fusion. And the third module is built to extensibly mine the interactions among multi-level attention maps by the proposed extensible co-attention fusion method. Extensive experimental results on three public datasets show the proposed model is significantly superior to the state-of-the-art methods, and demonstrate its effectiveness on two tasks of document-based and aspect-based MSA tasks.

Index Terms—Multimodal sentiment analysis, opinion mining, social analysis, multimodal fusion

1 Introduction

Multimodal sentiment analysis (MSA) aims to determine people's sentiment polarities towards topics or commodities from multimodal data such as texts and images. With the rapid development of social media and Internet technology, social network platforms such as Twitter[1] and Reddit[2] have become important ways for users to post or publish their content. The user-generated content in those platforms includes a variety of media forms such as texts, images and audios. Traditional sentiment analysis is to identify the users' sentiment polarity mainly from texts [1], [2]. However, with the increase of user-generated multimodal data, it is difficult to accurately recognize the users' sentiment only from textual content. MSA has been becoming increasingly attractive [3], and can be widely applied into many fields, including product marketing, public opinion monitoring and information recommendation [4], and so on [5].

At present, technically, MSA is mainly oriented to two issues: 1) video with acoustic, visual and textual modalities; 2) texts and images from user-generated content. The multimodal data parsed from videos are highly correlated to each other with spatially and temporally consistency. However, as for the texts and images from user-generated content, users may post texts and images separately, which results in that the multimodal data has not been aligned naturally for subsequent calculations. This brings challenges to multimodal sentiment analysis. This paper focuses on MSA tasks with texts and images from user-generated content.

Along this line, in the literature, some advanced MSA methods have been proposed to explore the correlations among multimodal data, including gate-based fusion methods [6], attention-based fusion methods [7], multi-task learning methods [8], and so on [9], [10], [11]. Typically, Kumar *et al*. [6] developed a gate-based method to flexibly coordinate unimodal and multimodal information flows, in which self-attention and gated mechanism are both used to learn multimodal fusion features. Wei *et al*. [7] proposed an attention-based fusion method to mine the consistent and complementary features with a cooperative network. Furthermore, Yu *et al*. [8] presented a multi-task learning framework with a group of unimodal subtasks and a joint multimodal task to explore the interaction information among multiple modalities. Moreover, Gkoumas *et al*. [11] designed a quantum probabilistic neural network to model the non-separability of multimodal data. For the above methods, performances on some public data sets show their usage for MSA tasks.

Technically, most of the previous works extract single-granularity feature from each modality. However, single-granularity features are insufficient to express the information

1. http://www.twitter.com/
2. https://www.reddit.com/

- *Xiaojun Xue, Chunxia Zhang, and Zhendong Niu are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail: {xiaojunx, cxzhang, zniu}@bit.edu.cn.*
- *Xindong Wu is with the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology, Hefei, Anhui 230009, China. E-mail: xwu@hfut.edu.cn.*

Manuscript received 20 July 2021; revised 30 Dec. 2021; accepted 12 Feb. 2022. Date of publication 1 Mar. 2022; date of current version 3 Apr. 2023.
The work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0104903 and in part by the National Natural Science Foundation of China under Grant 62072039.
(Corresponding author: Chunxia Zhang.)
Recommended for acceptance by L. Nie.
Digital Object Identifier no. 10.1109/TKDE.2022.3155290

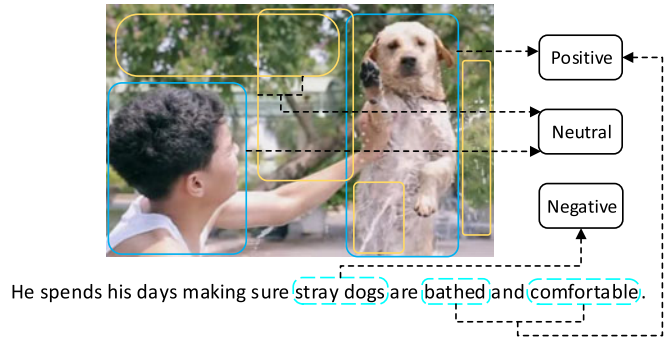


Fig. 1. Example demonstration of the importance of filtering, enhancing and fusing for multi-granularity features. The blue, yellow and cyan boxes represent image objects, image scenes and text word-level information, respectively.

contained in the raw multimodal data [6], [10], [11], [12]. Moreover, two or three types of multimodal features are considered to fuse together in most existing works [6], [7], [13]. Nevertheless, currently there does not exist an extensible framework that is able to fuse any number of different types of multi-granularity features from multimodal data so as to further improve the performance.

It has been observed that multi-granularity features can help grasp the local and non-local features [14]. Taking image as an example, the object features, scene features, and global features demonstrate three-level granularities. Their combination renders visual hints with different receptive fields from local regions to non-local regions, and thus helps to sufficiently mine the multi-scale visual information. In parallel, the data in the text modality can also be described with different scales with words and sentences, rending semantic information with different granularities.

However, there exist two typical kinds of noise in MSA tasks with multi-granularity features, which are explained as follows:

First, the features from different granularities may imply the inconsistent sentiment polarities. For example, in Fig. 1, the person and scenes indicate neutral sentiments, while different words represent positive or negative sentiments.

Second, the original features, which are extracted from the corresponding single-modalities by leveraging different pre-trained models, may contain some missing, redundant or even erroneous information. As a result, this may cause negative impacts on the multi-granularity feature fusion. As shown in Fig. 1, the original image scene features contain some noise to predict the user's sentiments. Thus, it is difficult to determine the sentiment polarities in the scenes with the yellow boxes in Fig. 1.

Although existing methods have made great progress in MSA tasks, most of them fulfilled the task under the assumption that the multimodal data is of high quality without noise. That is, most current methods ignore to deal with such noise before multimodal fusion [7], [10], [11]. Accordingly, their performances could degrade in complex situations where the multi-modalities are not well denoised and aligned consistently to each other.

According to the above analysis, the main challenges in MSA tasks under the framework of multi-granularity feature fusion can be summarized as the following two aspects: 1) insufficient multimodal feature extraction and the lack of feature filtering mechanism; 2) the complexity in information interaction between multi-granularity features and the lack of model extensibility for various types of multi-granularity features.

The above observations motivate us in this paper to develop multi-level attention maps to filter the above two types of noise so as to help enhance the representation ability for the original multi-granularity features.

Specifically, to address these challenges, we propose a multi-level attention map network (MAMN) for MSA tasks, which can extract adequately, filter and enhance effectively, and fuse extensibly the multi-granularity features of multimodal data. Fig. 2 shows the architecture of our designed model, where the network consists of three modules: 1) multi-granularity feature extraction module, 2) multi-level attention map generation module, and 3) attention map fusion module.

In MAMN model, the multi-granularity features are extracted from images and texts. That is, word-level features and sentence-level features are mined respectively from text data, while object features, scene features and image global features are extracted from image data. With these multi-granularity features, the MAMN is designed, trained and evaluated, rendering valuable observations and conclusions for the issues related to the MSA tasks.

The contributions of this paper are given as follows:

- A multi-level attention map network (MAMN) is proposed for multimodal sentiment analysis task, which helps to capture the consistent and heterogeneous correlations among multi-granularity features for multimodal sentiment analysis.
- Multi-level attention maps are designed to filter noise and enhance the representation ability for multi-granularity features before multimodal feature fusion. To the best of our knowledge, this is the first work on noise reduction and feature enhancement for multi-granularity features before the multimodal feature fusion.
- An extensible co-attention fusion method is developed to fuse various multi-level attention maps in an extensible and efficient way. This fusion mechanism has the ability to extract the complex interactions among multi-level attention maps, which helps to solve the extensibility and efficiency problem.
- Extensive experimental results on three public datasets indicate that the proposed approach outperforms significantly the state-of-the-art methods, and demonstrate its effectiveness and superiority on two tasks of document-based and aspect-based multimodal sentiment analysis.

The remainder of this paper is organized as follows. Section 2 discusses the related works. Section 3 explains the details of the proposed model. In Section 4, experimental results are reported and analyzed. Finally, the conclusion is summarized in Section 5.

2 Related Works

2.1 Unimodal Sentiment Analysis

Unimodal sentiment analysis methods are mainly developed for textual content. These methods can be classified

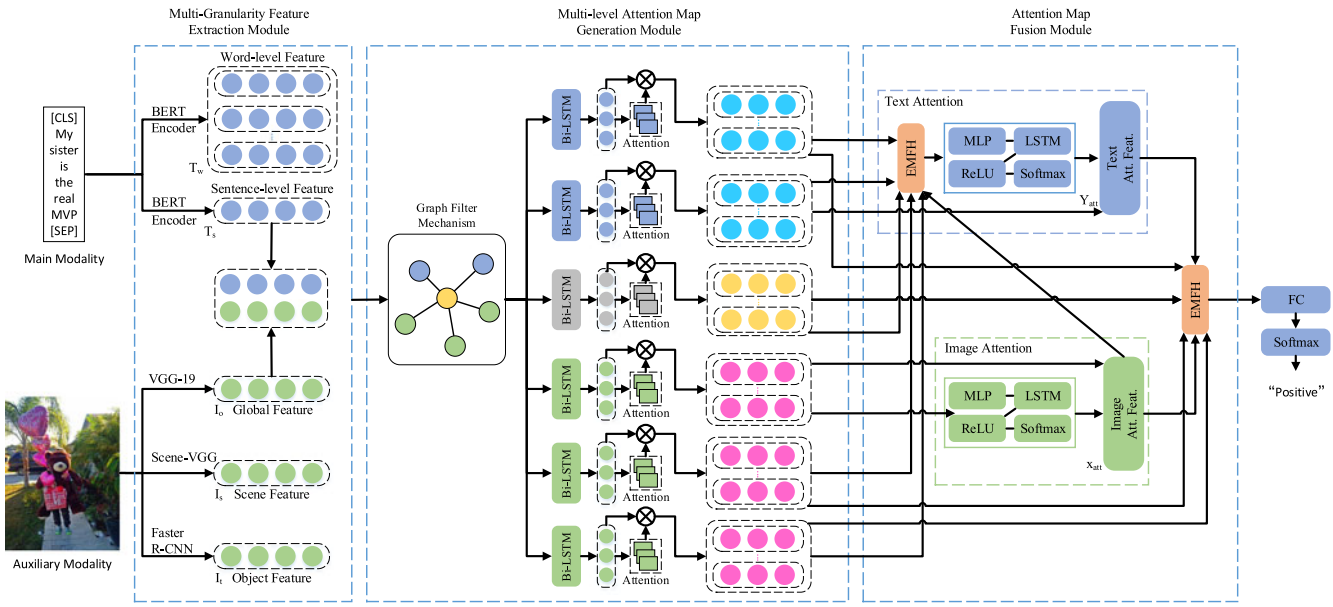


Fig. 2. Overview of multi-level attention map network for multimodal sentiment analysis.

into two groups: lexicon based approaches and machine learning based approaches [15]. As a typical lexicon based approach, SENTIWORDNET 3.0 automatically annotated all WORDNET synsets depending on their sentiment polarities, which included the semi-supervised learning step and the random-walk step [16]. For the machine learning based approaches, Thelwall *et al.* [15] proposed the SentiStrength algorithm optimized by machine learning to detect user behaviors and sentiment strength from textual content. In recent years, with the development of deep learning techniques, many works have adopted deep neural networks for sentiment analysis, such as those developed with convolutional neural network (CNN) [1] and recurrent neural network (RNN) [2] models, which demonstrates better performance than those of traditional methods.

However, technically, unimodal sentiment analysis methods are restricted to predicting users' sentiments from multimodal user-generated content.

2.2 Multimodal Sentiment Analysis

With the diversification of user-generated content, multimodal sentiment analysis has gradually become an important topic. In the literature, multimodal sentiment analysis approaches can be roughly divided into feature based methods and deep neural network (DNN) based methods. DNN based methods consist of feature-level fusion models, decision-level fusion models and hybrid fusion models. The following subsections give a brief about them for clarity.

2.2.1 Feature Based Methods

Early works on multimodal sentiment analysis adopted feature based methods to fuse information from different modalities [17], [18]. For instance, Rozgic *et al.* [17] addressed the sentence-level multimodal sentiment analysis and regarded that task as a multi-class classification problem. They classified emotions into multiple categories by leveraging a model based on the ensemble of support vector machines classifiers. Their model was verified by conducting five-way (happiness, neutral, sadness, anger and excitement) and four-way (happiness, neutral, sadness and anger) emotion recognition. In addition, Davidov *et al.* [18] used both Twitter tags and smileys as sentiment labels to enhance sentiment learning. They presented a novel supervised sentiment classification model, which was evaluated their model on data form Twitter. That model reduced the workloads of feature engineering.

However, for the feature based methods, the workloads required for feature engineering are time-consuming and laborious. Moreover, feature selection may cause the feature bias problem.

2.2.2 Deep Neural Network Based Methods

Without doubt, DNN based approaches have been received great attention in this issue. The methods in this family mainly differentiate from each other by the tricks on feature fusion. Accordingly, they can be further grouped into three sub-classes: methods with feature-level fusion, those with decision-level fusion, and those with hybrid fusion. By contrast, the hybrid fusion models fuse flexibly multimodal features, and there are no specific fusion boundaries in the fusion stage.

Feature-level fusion methods fuse the features extracted from different modalities and generate a fused feature to perform predictions. For instance, Hu *et al.* [19] utilized deep neural networks including Long Short-Term Memory (LSTM) [20] and inception model [21] to extract text features and image features, respectively. The dense layer was utilized to fuse those two types of features. They predicted latent sentiments of users from visual and textual information. Moreover, Gallo *et al.* [22] encoded the texts by convolution 1D layer, max-pooling layer and fully connected layer. The extracted text features were drawn over the image to enhance the information of the image and then CNNs were adopted to classify the sentiment. However, for these methods, the representation ability of fused features is insufficient, and the independence of each modality usually has an adverse effect on feature fusion.

In pipeline, the decision-level fusion methods first generate the prediction results or features of each modality, and then utilize fusion approaches such as voting and weighted summation to obtain the final sentiment results. Typically, Xu *et al.* [23] proposed a co-memory network model to mine the relations between images and texts. They presented the text-guided visual memory network and the image-guided textual memory network to leverage the visual and textual information to promote each other. Comparatively, the semantic information of images was leveraged to guide the sentiment words in texts by a vision guided attention module [14]. Object features and scene features were extracted from images to help extract the key words in sentences by using attention mechanism. The detailed semantic information of images was associated with the texts to perform multimodal emotion classification. The main disadvantage of the decision-level fusion methods is that they are restricted to mining the implicitly associated information in modality-specific features.

In the sub-class of hybrid fusion methods, many technical tricks have been employed, such as those via graph fusion [24], [25], gate-based fusion [6], attention-based fusion [7], [10], quantum-based fusion [11], [26], and fusions with multi-task learning [8], [27], and so on. The representative works of this type of methods are introduced as follows. Mai *et al.* [24] exploited a graph structure fusion method to fuse unimodal and multimodal features from different hierarchies. And before fusion, the distribution of multiple unimodal features were translated by adversarial learning. Kumar *et al.* [6] utilized gated mechanism to extensibly coordinate unimodal and multimodal information flows. Gated mechanism and self-attention were both used to selectively learn multimodal fusion features. Ghosal *et al.* [10] took into account the associated information between the target utterance and its neighboring utterances. They employed a novel multimodal attention network to mine the contextual information. Moreover, consistent and complementary features were definitely mined and distinguished by utilizing a cooperative network in the work of Wei *et al.* [7]. Chauhan *et al.* [27] took emotion analysis and sentiment analysis as subtasks to help detect multimodal sarcasm. A multi-task framework with two novel attention mechanisms was proposed to integrate multimodal features. Furthermore, Yu *et al.* [8] utilized three subtasks to generate single-modal labels, and jointly learned these three subtasks and a multimodal task to explore the interaction information among multiple modalities. Additionally, Gkoumas *et al.* [11] emphasized the indivisibility of multimodal information, and used a quantum probabilistic neural network to model the non-separability of multimodal data. Although the hybrid fusion methods fuse multimodal data based on different motivations and models, it is essentially mining consistent and different interaction correlations among multimodal features.

2.2.3 Summary of MSA

In summary, the current methods usually extract single-granularity features to perform multimodal feature fusion. Definitely, multi-granularity features can capture the different level (local and non-local) information within and across modalities. Nevertheless, multi-granularity features usually contain more noise, and the categories of the fused features have also increased a lot. As a result, the interactions among multi-granularity features are more complicated. Therefore, and more importantly, when utilizing multi-granularity features for fusion, how to effectively reduce noise and extensibly mine the interactions among multi-granularity features needs to be tackled.

Based on the above analysis, we propose a multi-level attention map network to solve those problems. The difference between our method and the existing methods is that our method focuses on how to filter noise and enhance the representation ability for multi-granularity features before feature fusion, and how to extensibly fuse various multi-level attention maps.

3 Proposed Model

3.1 Overview of the Proposed Framework

3.1.1 Problem Statement

Given a set T of multimodal tweets or reviews, each tweet or review t consists of a textual content $C = \{s_1, s_2, \ldots, s_m\}$ and an image set $I = \{i_1, i_2, \ldots, i_n\}$, where m is the number of sentences within the texts, and n is the number of images. The MSA task is to learn mapping functions f_α and f_β for document-based and aspect-based multimodal sentiment analysis, respectively. Here, function f_α takes the set of multimodal tweets as inputs, and predicts the sentiment label $l_r \in L_r$ of each tweet. In parallel, the function f_β takes the set of multimodal reviews as inputs, and predicts the sentiment label $l_a \in L_a$ of a given aspect for the review. In above, L_r is a set of sentiment polarities of tweets, and L_a is a set of sentiment polarities of a given aspect for the review. The functions f_α and f_β use the same model for learning. Formally, the MSA tasks can be formulated as follows:

$$l_r = f_\alpha(t, C, I), l_a = f_\beta(t, C, I). \tag{1}$$

3.1.2 Architecture of the Proposed Model

Fig. 2 shows the overall architecture of our multi-level attention map network (MAMN) for multimodal sentiment analysis. MAMN is composed of three modules: multi-granularity feature extraction module, multi-level attention map generation module, and attention map fusion module. The modalities of the input data are images and texts.

First, to sufficiently mine the multi-scale information contained in the multimodal data, multi-granularity features are extracted from both images and texts. Second, multi-level attention map is developed to reduce the noise in multi-granularity features. In this way, the representation ability of the original multi-granularity features is enhanced. Finally, aiming to fuse various multi-level attention maps and solve the extensibility and efficiency problem, an extensible co-attention fusion mechanism is proposed for multimodal feature fusion. In this way, the correlated information among the multi-granularity features can be extensibly and efficiently mined.

3.2 Multi-Granularity Feature Extraction Module

To sufficiently extract original features from different perspectives for multimodal data, multi-granularity features

are extracted from both texts and images. Different from single granularity features, multi-granularity features are capable of containing local and global information from different aspects, which enrich the feature descriptions possible with different levels of semantics.

3.2.1 Multi-Granularity Feature Extraction of Texts

Lexical features, especially those related to sentimental words within textual content, play an important role in sentiment analysis. In addition, the contextual semantic information on the sentence level offers useful clues for sentiment analysis. Hence, to capture local semantic features and global sentence contextual features, word-level features and sentence-level features are both extracted from texts to capture lexical and contextual sentiment characteristics. To this end, the pre-trained Bidirectional Encoder Representation from Transformers (BERT) [28] is employed to fulfill this job. The BERT [28] model can generate embeddings of words and sentences, which imply contextualized information of texts.

Specifically, given a sentence within a piece of textual content, the sum of hidden states of the last four layers in the BERT model is taken as the word-level feature $\mathbf{T}_w^o \in \mathbb{R}^{N\times b}$ of that piece of content. Here, N is the number of tokens or words in that sentence, b is the hidden size of BERT, the superscript o indicates unfiltered original feature. The $[CLS]$ representation of last layer in BERT hidden states is regarded as the sentence-level feature $\mathbf{T}_s^o \in \mathbb{R}^{1\times b}$.

3.2.2 Multi-Granularity Feature Extraction of Images

Using both the local and non-local information in images can help improve the representation ability of the visual features. Generally, the local information comes from the objects and the scenes in the image, while the non-local information is collected from the global feature of the image. Technically, the object features, scene features and global features yield three-level granularities. They are employed and fused together to describe the images. Their combination with different granularities grasps the visual contents with different receptive fields from local regions to non-local regions, and thus helps to sufficiently mine multi-scale visual information in the image.

Here, object features, scene features and global features are extracted to mine local and non-local traits in images. Specifically, the Faster R-CNN model [29] is utilized to extract the object features $\mathbf{V}_o^o$ of an image, the Scene-VGG model [30] is used to build the scene features $\mathbf{V}_s^o$ of that image, and the VGG-19 model [31] is employed to generate the global image features $\mathbf{V}_g^o$ of that image, where the superscript o indicates unfiltered original feature. Faster R-CNN utilizes Region Proposal Network to perform nearly cost-free region proposals by sharing convolutional features with the detection network. Scene-VGG model can handle the 365 scene classes classification problem and has been pre-trained on dataset-Place365 which contains millions of images. VGG-19 model utilizes very small convolution filters to build deeper networks at a low cost. Finally, those three kinds of features generated by Faster R-CNN, Scene-VGG and VGG-19 are transferred into same space and dimensionality by leveraging fully connected layers and ReLU functions

$$\mathbf{I}_o^o = ReLU(W_o(\mathbf{V}_o^o) + B_o), \tag{2}$$

$$\mathbf{I}_s^o = ReLU(W_s(\mathbf{V}_s^o) + B_s), \tag{3}$$

$$\mathbf{I}_g^o = ReLU(W_g(\mathbf{V}_g^o) + B_g). \tag{4}$$

3.3 Multi-Level Attention Map Generation Module

For the purpose of noise reduction and feature representation ability enhancement for multi-granularity features, the multi-level attention map generation module is designed. All the original multi-granularity features are filtered and enhanced before fusion. The correlated information within and between multi-granularity features are both introduced when filtering and enhancing the multi-granularity features, so as to assign high weights for the important parts in multi-granularity features. Technically, multi-level attention maps include the single-modality text attention maps, the single-modality image attention maps, and the multimodal attention map. Specifically, the single-modality text attention maps consist of text word-level attention map and text sentence-level attention. Single-modality image attention maps include the image global attention map, image object attention map, and image scene attention map. In addition, the multimodal attention map is generated from the text sentence-level feature and image global feature. As a result, the fine-grained features such as text word-level features and visual object features can be exploited to assist the fusion of global features extracted from texts and images.

3.3.1 Single-Modality Attention Maps Generation

Single-modality attention maps are generated to filter the noise and enhance the representation ability for all the multi-granularity features. This goal is achieved via learning mutually from the multi-granularity features to vote the important parts among the multimodal data. The correlated information within and among multi-granularity features are both introduced when filtering two kinds of noise and enhancing the multi-granularity features. Technically, we generate single-modality attention maps respectively for text and image.

First, a graph filter mechanism is introduced to deal with the first type of noise that the original multi-granularity features may contain some missing, redundant or even erroneous information. Specifically, the graph is constructed as follows. The text sentence level feature and image global feature are concatenated to generate original multimodal feature $\mathbf{F}_m^o$. The central node of the graph is defined on the generated multimodal feature, and one-hop neighbors of that central node are defined on multi-granularity features. Moreover, the edges denote the categories of multi-granularity features (such as text word-level category and image object category). The reason for this treatment lies in that the categories and their corresponding feature vectors can be taken as attributes and attribute values of the raw multimodal data. Furthermore, by introducing the neighboring nodes and edges information of each node, different nodes can be assigned different weights via learning mutually from the multi-granularity features. In this way, the representations of all nodes are updated. Correspondingly, with

the weighted average, the noise between multi-granularity features is largely reduced.

Inspired by graph attention network (GAT) [32], but different from it, both the categories (edges) and the feature values (nodes) of multi-granularity features are both introduced to filter noise in this subsection. The inputs of attention map generation module are category matrix $D \in \mathbb{R}^{5\times T}$ and value matrix $E \in \mathbb{R}^{6\times P}$, where T is the dimension of each category embedding, and P stands for that of each value. The category matrix D includes five categories of the multi-granularity features, i.e., text sentence-level, text word-level, image global, image object and image scene, which are randomly initialized and then learned during the training process. Value matrix E collects together these five multi-granularity features and a multimodal feature e_m. Here, e_m is obtained by concatenating the text sentence-level feature and the image global feature along the first dimension. That multimodal feature is used as the central node in the graph.

For the purpose of generating filtered features for all nodes, a representation c_{ijk} is learned for every category-value pair, as shown in Eq. (5)

$$c_{ijk} = \mathbf{W}_{ijk}[e_i||e_j||d_k], \tag{5}$$

where e_i and e_j are the ith and jth rows of value matrix E, d_k denotes the kth row of category matrix D, $||$ stands for the concatenation operation, and $\mathbf{W}_{ijk}$ is the linear transformation matrix.

To filter and enhance the original features, attention mechanism is utilized to assign different weights. The attention weights are calculated as follows:

$$\alpha_{ijk} = \frac{exp(c_{ijk})}{\sum_{n\in N_i}\sum_{r\in R_{in}} exp(c_{inr})}, \tag{6}$$

where N_i represents the neighborhood of node e_i, and R_{in} denotes the edge connecting nodes e_i and e_n. The filtered feature of the node e_i is the sum of each pair representation weighted by their attention weights

$$h_i = \sum_{j\in N_i}\sum_{k\in R_{ij}} \alpha_{ijk} c_{ijk}. \tag{7}$$

By performing the above calculation on all nodes, we can get five new multi-granularity features $\mathbf{T}_w$, $\mathbf{T}_s$, $\mathbf{I}_o$, $\mathbf{I}_s$, $\mathbf{I}_g$ and one new multimodal feature $\mathbf{F}_m$, in which the noise among multi-granularity features are filtered. Here, these five new multi-granularity features are denoted as the weighted multi-granularity features.

Next, the "Bi-LSTM + attention mechanism" is developed to deal with the second type of noise that the original extracted features may be independent, one-sided, and not accurate enough.

Technically, Bi-LSTM is employed to mine the long dependencies of different parts of the feature matrices transformed from the weighted multi-granularity features. By contrast to the traditional LSTM, Bi-LSTM is performed more stable to grasp the dependency among the long sequences with two directions. In this way, the error aggregation and the noise can be eased to some degree with the support of the context information hidden in the sequence.

The attention mechanism is developed to learn the weights from the transformed multi-granularity features to vote the important parts of the features, which further helps enhance the representation ability of the features. Thus, the joint of the "Bi-LSTM + attention mechanism" makes the feature learning more robust.

The single-modality attention maps are generated for all the multi-granularity features. In other words, text word-level attention map, text sentence-level attention map, image global attention map, image object attention map and image scene attention map are generated to sufficiently filter and enhance the features extracted from multimodal data. Here, we take the text sentence-level attention map generation as an example to explain the details.

To this end, the weighted text sentence-level feature $\mathbf{T}_s$ is first transformed into a feature matrix by reshaping operation, and then fed into Bi-LSTM to obtain the text sentence-level hidden state h_s. Then the attention mechanism is applied over h_s to obtain the weights $\alpha_{s,i}$ for different parts in the text sentence-level hidden state h_s. We quantify the importance of different parts in the text sentence-level hidden state h_s as the similarity of $h_{s,i}$ with the text sentence-level query vector q_s

$$h_s = Bi - LSTM(\mathbf{T}_s), \tag{8}$$

$$\alpha_{s,i} = \frac{exp(h_{s,i}q_s)}{\sum_{j=1}^{L} exp(h_{s,j}q_s)}, \tag{9}$$

where q_s represents the text sentence-level query vector, which is randomly initialized and learned during training, In this work, six distinct query vectors are introduced for the generation of multi-level attention maps, Then, different attention weights are assigned to different parts in multi-granularity features, which can effectively filter the noise in multi-granularity features. For clarity, L denotes the number of different parts in the text sentence-level hidden state h_s. That is, L is equal to the number of rows in the weighted text sentence-level feature $\mathbf{T}_s$. Then, the different parts $h_{s,i}$ in text sentence-level hidden state are weighted and summed according to its attention weights $\alpha_{s,i}$. Finally, the text sentence-level attention map $\mathbf{M}_s$ is calculated as follows:

$$\mathbf{M}_s = \sum_{i=1}^{L} \alpha_{s,i}\mathbf{h}_{s,i}. \tag{10}$$

Finally, it is worthy pointing out that the generations for text word-level attention map $\mathbf{M}_w$, image global attention map $\mathbf{M}_g$, image object attention map $\mathbf{M}_o$ and image scene attention map $\mathbf{M}_c$ follow the same steps for text sentence-level attention map $\mathbf{M}_s$. In this way, they are generated to reduce the noise and enhance the representation ability for the original multi-granularity features.

3.3.2 Multimodal Attention Map Generation

Multimodal attention map is generated to filter those two kinds of noise in multimodal feature. In this work, fine-grained features such as text word-level features and image object features are utilized to provide supplementary information in the fusion stage, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2022.3155290. Therefore, multimodal attention map is generated from global features of text and

image, i.e., text sentence-level feature and image global feature.

As for the original multimodal feature $\mathbf{F}_m^o$, the first kind of noise has been filtered by utilizing graph filter mechanism in Section 3.3.1. Then, similar to the calculation process of single-modality attention maps, Bi-LSTM and attention mechanism are utilized to reduce the second kind of noise and enhance representation ability for the multimodal features. The multimodal hidden state h_m is obtained by feeding the weighted multimodal feature $\mathbf{F}_m$ into Bi-LSTM

$$h_m = Bi - LSTM(\mathbf{F}_m). \tag{11}$$

Next, the attention mechanism is utilized to generate multimodal attention weights $\beta_{m,i}$. The importance of different parts $h_{m,i}$ in multimodal hidden state h_m is quantified as the similarity of $h_{m,i}$ with the multimodal query vector q_m. The multimodal attention map $\mathbf{M}_a$ is obtained by weighting and summing different parts $h_{m,i}$ in the multimodal hidden state according to its attention weight $\beta_{m,i}$, which is calculated as follows:

$$\beta_{m,i} = \frac{exp(h_{m,i}q_m)}{\sum_{j=1}^{K} exp(h_{m,j}q_m)}, \tag{12}$$

$$\mathbf{M}_a = \sum_{i=1}^{K} \beta_{m,i}\mathbf{h}_{m,i}. \tag{13}$$

where q_m represents the multimodal query vector, which is randomly initialized and learned during training. In Eq. (12), K denotes the number of different parts in the multimodal hidden state h_m.

3.4 Attention Map Fusion Module

Aiming to fuse various multi-level attention maps and solve the extensibility and efficiency problem, an extensible co-attention fusion mechanism is proposed for multimodal feature fusion. In this way, the correlated information among the multi-granularity features could be extensibly and efficiently mined.

Technically, most previous works only fused at most three multimodal features, and the model complexity and computational cost are greatly increased in the case that more features are fused together. Therefore, extensible co-attention fusion method is developed to tackle this problem. Specifically, extensible co-attention fusion method incudes two mechanisms: extensible multimodal factorized high-order pooling mechanism (EMFH) and modality-specific attention mechanisms. Those two mechanisms are designed to explore the consistency and distinguish the heterogeneity for multi-level attention maps, respectively.

This fusion mechanism can allow to fuse any number of modalities, and has the ability to extract the complex interaction among multi-level attention maps, which could help to solve the extensibility and efficiency problem.

3.4.1 Extensible Multimodal Factorized High-Order Pooling Mechanism

Extensible multimodal factorized high-order pooling mechanism (EMFH) is proposed to explore the consistency of the multi-level attention maps. In attention map fusion module, there are six attention maps, namely, $\mathbf{M}_s$, $\mathbf{M}_w$, $\mathbf{M}_g$, $\mathbf{M}_o$, $\mathbf{M}_c$, $\mathbf{M}_a$ from different levels that need to be fused to explore the interactions between them. Previous works only fused two or three multimodal features, and the model complexity and computational cost are greatly increased when various multimodal features are fused. In this section, we focus on effectively and efficiently fusing various multimodal features.

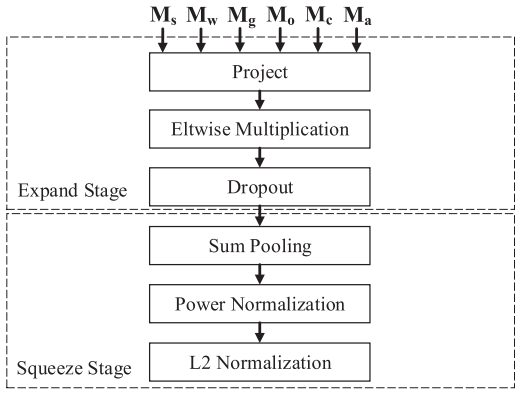


Fig. 3. Architecture of extensible multimodal factorized bilinear pooling (EMFB) block.

By contrast, the multimodal factorized high-order pooling mechanism [33] was originally used in visual question answering task to fuse image features and text features. However, it can only handle two kinds of feature inputs. In this work, the number of multi-level attention maps that need to be fused is more than two, so EMFH mechanism is proposed to effectively and efficiently fuse those attention maps. EMFH is formed by cascading multiple extensible multimodal factorized bilinear (EMFB) pooling blocks. The architecture of EMFB block is shown in Fig. 3. Each EMFB consists of expand stage and squeeze stage. Specifically, in EMFB, fully connected layers are used to project the inputs into same space and dimensionality. Elementwise multiplication and pooling layers are utilized to integrate features and explore the consistency of multi-level attention maps. A dropout layer is leveraged to prevent overfitting. Power normalization and L_2 normalization layers are added to prevent the model from converging to an unsatisfactory local minimum.

In the previous work, the earliest bilinear multimodal fusion model is defined as follows:

$$\mathrm{z}_i = x^T\mathbf{W}_i y, \tag{14}$$

where x is the image feature, y is the text feature, $\mathbf{W}_i$ is a projection matrix, z_i is the output of the bilinear model. The trained matrix is $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_o]$ and then we can get the o-dimensional output $\mathrm{z} = [\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_o]$.

According to the matrix factorization methods[34], [35], [36], [37], [38] for unimodal data, we can get two low-rank matrices after factorization of the projection matrix $\mathbf{W}_i$

$$z_i = x^T\mathbf{U}_i{\mathbf{V}_i}^T y = \sum_{d=1}^{k} x^T u_d {v_d}^T y = I^T({\mathbf{U}_i}^T x \circ {\mathbf{V}_i}^T y), \tag{15}$$

where $\mathbf{U}_i$ and $\mathbf{V}_i$ are the factorized learnable parameter matrices, k is the latent dimensionality of the factorized

matrices $\mathbf{U}_i$ and $\mathbf{V}_i$, I denotes an all-one vector, and $\circ$ represents the Hadamard product of two feature vectors.

Next, the factorized matrices $\mathbf{U}_i$ and $\mathbf{V}_i$ form the learned weights $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_o]$ and $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_o]$. These three-order matrices $\mathbf{U}$ and $\mathbf{V}$ can be reformulated as two-order matrices $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ by reshaping operations.

To enable the fusion module to handle multi-level attention maps, we extend traditional multimodal factorized bilinear pooling mechanism as follows:

$$\begin{aligned} z_{exp} &= \mathrm{EMFB}_{\exp}(\mathbf{M}_s, \mathbf{M}_w, \mathbf{M}_g, \mathbf{M}_o, \mathbf{M}_c, \mathbf{M}_a) \\ &= Dropout(tanh(\tilde{\mathbf{U}}_s^T \mathbf{M}_s \circ \tilde{\mathbf{U}}_w^T \mathbf{M}_w \circ \tilde{\mathbf{U}}_g^T \mathbf{M}_g \\ &\quad \circ \tilde{\mathbf{U}}_o^T \mathbf{M}_o \circ \tilde{\mathbf{U}}_c^T \mathbf{M}_c \circ \tilde{\mathbf{U}}_a^T \mathbf{M}_a)), \end{aligned} \tag{16}$$

$$\begin{aligned} z &= \mathrm{EMFB}_{sqz}(z_{\exp}) \\ &= Norm(SumPool(z_{\exp})), \end{aligned} \tag{17}$$

where $\mathbf{M}_s$, $\mathbf{M}_w$, $\mathbf{M}_g$, $\mathbf{M}_o$, $\mathbf{M}_c$ and $\mathbf{M}_a$ denote six distinct multi-level attentions maps in Section 3.3. Aiming at obtaining the output feature z of EMFB, different from previous works, the learned weights are six two-order matrices $\tilde{\mathbf{U}}_s^T$, $\tilde{\mathbf{U}}_w^T$, $\tilde{\mathbf{U}}_g^T$, $\tilde{\mathbf{U}}_o^T$, $\tilde{\mathbf{U}}_{\mathbf{c}}^{\mathbf{T}}$ and $\tilde{\mathbf{U}}_a^T$, corresponding to six multi-level attentions. $\circ$ denotes the Hadamard product. exp and sqz represent the expand stage and squeeze stage in EMFB. z_{exp} is the output of expand stage in EMFB, which is also the input of squeeze stage. z is the output feature of EMFB. $Dropout$ represents the dropout layer. $SumPool$ represents the sum pooling layer, and $Norm$ represents the normalization layer.

The EMFB method can fuse any kinds of input features by changing the number of elements for the Hadamard product. The proposed fusion mechanism has good expansibility and can fuse any number of multimodal features.

Multiple EMFB blocks are cascaded to form the EMFH mechanism, which is used to extract more complex interactions between multi-level attention maps. To cascade p EMFB blocks, the z_{exp} can be modified as follows:

$$\begin{aligned} z_{exp}^i &= \mathrm{EMFB}_{exp}^i(\mathbf{M}_s, \mathbf{M}_w, \mathbf{M}_g, \mathbf{M}_o, \mathbf{M}_c, \mathbf{M}_a) \\ &= z_{exp}^{i-1} \circ (Dropout(tanh(\tilde{\mathbf{U}}_s^{i^T} \mathbf{M}_s \circ \tilde{\mathbf{U}}_w^{i^T} \mathbf{M}_w \\ &\quad \circ \tilde{\mathbf{U}}_g^{i^T} \mathbf{M}_g \circ \tilde{\mathbf{U}}_o^{i^T} \mathbf{M}_o \circ \tilde{\mathbf{U}}_c^{i^T} \mathbf{M}_c \circ \tilde{\mathbf{U}}_a^{i^T} \mathbf{M}_a))). \end{aligned} \tag{18}$$

where i denotes the ith EMFB blocks, z_{exp}^i can be calculated according to z_{exp}^{i-1}. Then the output z_i of the ith EMFB block can be computed by Eq. (15). Here, z_{exp}^0 is an all-one vector. It is noted that any item in the inputs of EMFB can be replaced with an attention feature in attention map fusion module.

Next, the output features of p EMFB blocks are concatenated to obtain the final output z of the EMFH^p mechanism

$$z = \mathrm{EMFH}^p = [z^1, z^2, \ldots, z^p]. \tag{19}$$

3.4.2 Extensible Co-Attention Fusion

Extensible co-attention fusion method (ECAF) is designed to fuse various multi-level attention maps in an extensible and efficient way, which can explore the consistency and distinguish the heterogeneity for multimodal data.

Technically, the extensible co-attention fusion method consists of the proposed extensible multimodal factorized high-order pooling mechanism (EMFH), text attention and image attention. These two modality-specific attention mechanisms are developed to distinguish the heterogeneity for multimodal data. We argue that in a specific scenario, a certain modality contributes more to the overall sentiment analysis. Therefore, text modality is taken as the main modality for emotion judgment when fusing the multi-level attention maps. The EMFH mechanism is used twice in this section, and the EMFH mechanism is first used in the attention module corresponding to the main modality.

In the attention map fusion module, the text sentence-level attention map and the image global attention map are regarded as the main information source, and other multi-level attention maps assist them to perform multimodal feature fusion. As shown in Fig. 2, image attention is composed of multilayer perceptron (MLP), LSTM layer, ReLU activation function and attention mechanism. Text attention has the similar structure to image attention, however, the difference is that EMFH mechanism is applied to fuse multi-level attention maps and image attention feature.

First, image attention module takes the image global attention map $\mathbf{M}_g$ as input to mine the deep features of images, and generate the image attention feature $\mathbf{X}_{att}$

$$\mathbf{X}_{att} = ImageAttention(\mathbf{M}_g). \tag{20}$$

Second, EMFH mechanism is utilized to explore the correlations between image attention features and multi-level attention maps. The image attention feature $\mathbf{X}_{att}$, text sentence-level attention map $\mathbf{M}_s$, text word-level attention map $\mathbf{M}_w$, image object attention map $\mathbf{M}_o$, image scene attention map $\mathbf{M}_c$ and multimodal attention map $\mathbf{M}_a$ are fed into EMFH within text attention to generate the fusion feature $\mathbf{Y}_f$

$$\mathbf{Y}_f = EMFH(\mathbf{X}_{att}, \mathbf{M}_s, \mathbf{M}_w, \mathbf{M}_o, \mathbf{M}_c, \mathbf{M}_a). \tag{21}$$

Third, MLP, LSTM layer and ReLU activation function are utilized to process the fusion feature $\mathbf{Y}_f$, which can mine deep information of the fuse feature $\mathbf{Y}_f$. Attention mechanism is utilized to weight the text sentence-level attention map within the main modality, and then text attention feature $\mathbf{Y}_{att}$ is built by leveraging the above text attention

$$\mathbf{Y}_{att} = TextAttention(\mathbf{Y}_f). \tag{22}$$

Next, EMFH mechanism is used again to explore the correlations between multimodal attention features and multi-level attention maps. Specifically, the text attention feature $\mathbf{Y}_{att}$, the image attention feature $\mathbf{X}_{att}$ and the auxiliary multi-level attention maps $\mathbf{M}_w$, $\mathbf{M}_o$, $\mathbf{M}_c$, $\mathbf{M}_a$ are fed into another EMFH to get the fusion feature v_m used for sentiment classification

$$v_m = EMFH(\mathbf{Y}_{att}, \mathbf{X}_{att}, \mathbf{M}_w, \mathbf{M}_o, \mathbf{M}_c, \mathbf{M}_a). \tag{23}$$

Finally, the fusion feature v_m generated in the attention map fusion module is input into fully connected layer and softmax layer to obtain the prediction result l_r (or l_a). The multi-class cross entropy loss function is used as the loss

function for this model

$$l_r(or\ l_a) = softmax(FC(v_m)). \tag{24}$$

Algorithm 1 shows the flow of the multi-level attention map network for multimodal sentiment analysis. The inputs contain textual content C and image set I. First, multi-granularity features are extracted from images and texts (lines 1 and 2). Next, those extracted multi-granularity features of texts and images are utilized to generate multi-level attention maps (lines 4-7). Then, the multi-level attention maps are fused by using the extensible co-attention fusion method (lines 8-11). Finally, the multimodal sentiment classification results are generated through the fully connected layer and the softmax layer (line 12).

4 Experiments

4.1 Datasets

The proposed model is evaluated on three public multimodal sentiment analysis datasets: MVSA-Single, MVSA-Multi [39] and Multi-ZOL [13]. MVSA-Single dataset and MVSA-Multi dataset[3] come from the work of Niu *et al.* [39], which are derived from Twitter and consist of text-image pairs. There are 5,129 samples labeled by one person in MVSA-Single dataset and 19,600 samples labeled by three persons in MVSA-Multi dataset. The preprocessing about sentiment labels for those two datasets is the same as the work in Xu *et al.* [14].

In the MVSA-Single dataset, texts and images have positive, neutral and negative sentiment labels. Each sample is annotated by one annotator. For each sample, what we need to do is to generate the sample label based on its text label and image label. When the text label and the image label are identical, the sample label is the text (or image) label. Some samples have inconsistent text label and image label. In order to ensure the availability of the data, as in the previous works [14], [40], [41], we remove samples where one of the text and image labels is positive and the other is negative. When one label is positive (or negative) and the other label is neutral, we regard the sample label as positive (or negative). Finally, 4,511 multimodal samples are obtained.

For the MVSA-Multi dataset, texts and images also have positive, neutral, and negative emotional labels. The difference is that each sample is annotated by three annotators. First, we vote for multiple sentiment labels of texts and images. When at least two texts or images have the same labels, the sample is considered valid. After voting, there is only one label for each text and each image. Next, as with the processing of MVSA-Single dataset, the samples with inconsistent text labels and image labels are deleted. Finally 17024 multimodal samples are obtained.

The Multi-ZOL dataset[4] is built by Xu *et al.* [13] and used for aspect-level multimodal sentiment analysis, including 5,288 samples of multimodal mobile phones reviews. The dataset is derived from ZOL.com. Each sample of multimodal reviews comprises a text review, an image set and 1-6 aspects. The six aspects contain performance configuration, price-performance ratio, appearance and feeling, battery life, photographing effect, and screen of mobile phones. Each aspect of each review is given an integer score from 1 to 10. The score is treated as the sentiment label in our experiment as the same with previous work [13]. Each aspect is paired with multimodal review and 28,469 aspect-review pairs of samples are obtained.

Algorithm 1. Multimodal Sentiment Analysis Based on a Multi-Level Attention Map Network

Input: textual content C, image set I,
Output: Sentiment classification results l_r or l_a

1 Extract original global feature $\mathbf{I}_g^o$, object feature $\mathbf{I}_o^o$ and scene feature $\mathbf{I}_s^o$ for images;
2 Extract original sentence-level features $\mathbf{T}_s^o$ and word-level features $\mathbf{T}_w^o$ for text;
3 **for** $e \leftarrow 1$ **to** $Epochs$ **do**
4 $\mathbf{F_m^o} \leftarrow (\mathbf{T_s^o}, \mathbf{I_g^o})$ by concatenation;
5 $\mathbf{T_s}, \mathbf{T_w}, \mathbf{I_g}, \mathbf{I_s}, \mathbf{I_o}, \mathbf{F_m} \leftarrow (\mathbf{T_s^o}, \mathbf{T_w^o}, \mathbf{I_g^o}, \mathbf{I_s^o}, \mathbf{I_o^o}, \mathbf{F_m^o})$ by Eqs. (5), (6) and (7);
6 $\mathbf{M_s}, \mathbf{M_w}, \mathbf{M_g}, \mathbf{M_c}, \mathbf{M_o} \leftarrow (\mathbf{T_s}, \mathbf{T_w}, \mathbf{I_g}, \mathbf{I_s}, \mathbf{I_o})$ by Eqs. (8), (9) and (10);
7 $\mathbf{M_a} \leftarrow \mathbf{F_m}$ by Eqs. (11), (12) and (13);
8 $\mathrm{X}_{att} \leftarrow \mathbf{M_g}$ by Eq. (20);
9 $\mathrm{Y}_f \leftarrow (\mathrm{X}_{att}, \mathbf{M_s}, \mathbf{M_w}, \mathbf{M_o}, \mathbf{M_c}, \mathbf{M_a})$ by Eq. (21);
10 $\mathrm{Y}_{att} \leftarrow \mathrm{Y}_f$ by Eq. (22);
11 $v_m \leftarrow (\mathbf{Y_{att}}, \mathbf{X_{att}}, \mathbf{M_w}, \mathbf{M_o}, \mathbf{M_c}, \mathbf{M_a})$ by Eq. (23);
12 $l_r(or\ l_a) \leftarrow v_m$ by FC and softmax;
13 Calculate the loss and perform back propagation;
14 **end**

All three datasets are randomly divided into the training set, the validation set, and the test set with split ratio of 8:1:1. The batch size, learning rate, dropout probability, output vector length of fusion modules, the number of fusion modules, the number of epochs are equal to 128, 0.0001, 0.1, 800, 4, 200, respectively. The hidden size and the number of layers for Bi-LSTM is 512 and 2, respectively. And the hidden size of LSTM is set to 512. All experiments are conducted under the Ubuntu system with 3,090 GPU. As for the fusion method, the time complexity of our MAMN model is about $O(log\, n)$, while the time complexity of the NMCL model [7] scales in $O(log\, n^2)$. Because all features only need to be fused once by leveraging factorization fusion in the EMFH mechanism, while every two kinds of features are interacted in the NMCL model. Moreover, the parameter amount of the proposed MAMN model is about 1.83×10^8. By contrast, the parameter amount of the NMCL model [7] is about 1.27×10^8. Although our model has more parameters, it can filter the noise and extensibly fuse various multi-granularity features, and has achieved a significant performance improvement.

4.2 Experimental Results

The evaluation metrics used in the experiment are accuracy and F1-measure. For the MVSA-Single and MVSA-Multi datasets, we have compared the performance of the proposed model with 19 baseline methods:

1) *SentiBank* [42] can mine 1,200 Adjective Noun Pairs (ANP) from every image to perform sentiment classification.

3. http://www.mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
4. https://github.com/xunan0812/MIMN

2) *Scene, Object and Scene+Object* [14] are three variants of MultiSentiNet, which only utilize image information.
3) *SentiStrength* [15] detects user behaviors and sentiment strength from textual English content.
4) *CNN-Multichannel* [1] introduces a multichannel convolutional neural networks with multiple filter widths.
5) *LSTM-Avg* [2] applies Long Short-term Memory (LSTM) networks to sentiment classification task.
6) *SentiBank+SentiStrength* [42] combines SentiBank and SentiStrength to perform sentiment classification.
7) *CNN-Multi* [9] uses multiple CNNs to extract multimodal features and capture the interactions.
8) *DNN-LR* [12] uses multiple deep convolutional neural network (DNN) to extract features for texts and images.
9) *CBOW+DA+LR* [43] inputs the information of one text windows into the CBOW model at a time. The corresponding image is processed by denoising autoencoder.
10) *HSAN* [44] generates image captions to supplement the textual information. A hierarchical structure network based on LSTM and attention mechanism are utilized for features fusion.
11) *MultiSentiNet* [14] leverages the semantic information of images to guide the sentiment words in texts. Object features and scene features are extracted from images.
12) *MMMU-BA* [10] considers the associated information between the target utterance and its neighboring utterances. A novel multimodal attention network is employed to mine the contextual information.
13) *Co-Memory Network (CoMN-Hop4 and CoMN-Hop6)* [23] mines the relations between images and texts introduces by introducing a co-memory network.
14) *CFF-ATT* [41] performs multimodal fusion based on symmetry and attention mechanism.
15) *NMCL* [7] distinguishes the consistent and complementary features by using a cooperative network.
16) *TBJE* [45] uses distinct transformer-based encoders, and combines attention mechanism and residual connection.
17) *MTLF* [27] detects multimodal sarcasm by taking emotion analysis and sentiment analysis as subtasks. A multi-task framework with two novel attention mechanisms was proposed to integrate multimodal features.
18) *Self-MM* [8] jointly learns three subtasks and a multimodal task to explore the interaction information among multiple modalities.
19) *MMIM* [46] combines multimodal sentiment analysis and mutual information to prevent the loss of task-related information.

As for the Multi-ZOL dataset, 13 baseline methods are compared with our model:

1) *LSTM* [47] utilizes multiple Long Short-term Memory (LSTM) networks to learn the feature of textual content.
2) *MemNet* [48] inputs word embeddings into multiple computational layers based on attention mechanism to mine the deep features of texts.
3) *ATAE-LSTM* [47] introduces the aspect information. Attention mechanism and aspect embedding are combined with LSTM in this model.
4) *IAN* [49] uses two LSTM networks to model target and context. Interactive attention is proposed to interactively learn the representations of target and context.
5) *RAM* [50] builds a memory model based on bidirectional LSTM. For each target, position-weighted memory is proposed to generate a tailor-made input memory.
6) *Co-Memory+Aspect* [23] captures the interactive information by introducing multiple memory hops. Visual memory network and textual memory network are proposed to mine features of images and texts, respectively.
7) *MIMN* [13] inputs text embeddings, aspect embeddings and image features into to multi-interactive memory networks to capture the interaction between multimodal features and the self influences in each modality feature.

The model structures of MMMU-BA [10], NMCL [7], TBJE [45], MTLF [27], MMIM [46] and Self-MM [8] are the same as that in MVSA-Single and MVSA-Multi datasets except the inputs and prediction labels.

The performance comparison result on the MVSA-Single and MVSA-Multi datasets is shown in Table 1. And Table 2 shows the comparison result on the Multi-ZOL dataset. We separately list the comparison results and employ different

TABLE 1
Comparative Results of MAMN and Baselines on MVSA-Single Dataset and MVSA-Multi Dataset

Datasets	MVSA-Single		MVSA-Multi	
Method	Acc	F1	Acc	F1
SentiBank	45.22	43.80	55.02	51.15
Scene	63.64	60.40	67.69	65.24
Object	62.08	56.45	65.80	64.75
Scene+Object	64.08	62.33	67.98	66.23
SentiStrength	49.86	48.45	50.57	49.84
CNN-Multichannel	65.19	62.55	65.57	63.24
LSTM-Avg	65.85	64.11	65.69	65.63
SentiBank+SentiStrength	52.05	50.08	65.62	55.36
CBOW+DA+LR	63.86	63.52	64.22	63.73
CNN-Multi	61.20	58.37	66.39	64.19
DNN-LR	61.42	61.03	67.86	66.33
HSAN	66.83	66.90	68.16	67.76
MultiSentiNet	69.84	69.63	68.86	68.11
MMMU-BA	68.72	68.35	69.24	68.76
MTLF	68.84	68.51	69.78	69.39
CoMN-Hop4	69.18	68.29	69.92	69.83
CoMN-Hop6	70.51	70.01	68.92	68.83
CFF-ATT	71.44	71.06	69.62	69.35
TBJE	72.62	72.21	71.50	71.13
NMCL	71.72	71.39	72.38	72.02
MMIM	73.26	72.83	74.54	74.13
Self-MM	72.37	71.96	75.19	74.88
MAMN w/o AM	71.85	71.44	73.61	73.02
MAMN w/o ECAF	67.21	66.58	69.52	68.94
MAMN-Glove	75.86	75.19	77.47	76.78
MAMN	**76.57**	**76.08**	**78.34**	**77.92**

TABLE 2
Comparative Results of MAMN and Baselines on Multi-ZOL Dataset

Method	Accuracy	Macro-F1
LSTM	58.92	57.29
MemNet	59.51	58.73
ATAE-LSTM	59.58	58.95
IAN	60.08	59.47
RAM	60.18	59.68
Co-Memory+Aspect	60.43	59.74
MTLF	60.63	60.25
MMMU-BA	60.84	60.57
MIMN	61.59	60.51
TBJE	62.35	61.83
MMIM	62.87	62.31
NMCL	63.65	63.22
Self-MM	63.93	63.45
MAMN w/o AM	71.18	70.69
MAMN w/o ECAF	61.47	61.04
MAMN-Glove	73.86	73.02
MAMN	**75.41**	**74.79**

baselines for MVSA datasets and Multi-ZOL dataset. Because MVSA datasets is built to solve the document-based sentiment analysis task, while Multi-ZOL dataset is designed to tackle the aspect-based multimodal sentiment analysis task. Those two kinds of sentiment analysis tasks have different intensions. Specifically, MVSA datasets is used to evaluate the sentiment polarities in tweets, including three polarities: positive, neutral and negative. Differently, Multi-ZOL dataset is originally constructed to evaluate the sentiment scores of given aspects for the reviews. Each aspect of a review is given an integer score from 1 to 10. To sufficiently and fairly compare with the baseline methods, the experimental results of related works come from the previous works, which solves these two tasks separately.

The first group (SentiBank, Scene, Object and Scene+Object methods) in Table 1 shows the performance of different methods only using image data of MVSA datasets. The second group (SentiStrength, CNN-Multichannel and LSTM-Avg methods) demonstrates different methods using text data in the MVSA datasets. The third group (the rest of the methods in Table 1) shows the performance of different methods using multimodal data of MVSA datasets. In Table 2, LSTM, MemNet, AEAT-LSTM, IAN, RAM use the textual data in the dataset, while the other models utilize multimodal data in Multi-ZOL dataset.

We reproduced MMMU-BA [10], NMCL [7], TBJE [45], MTLF [27], MMIM [46] and Self-MM [8] models and test them on those three datasets. The results of the rest of baseline methods are from previous works [14], [23], [41]. To make a fair comparison, as in the previous work, Glove [51] is also used to generate word embeddings in the sentences for all three datasets. "MAMN w/o AM" and "MAMN w/o ECAF" are two variants of our MAMN model, which are reported to help analyze the effect of different modules. "MAMN w/o AM" denotes replacing the multi-level attention maps with original multimodal features without feature filtering and enhancing. "MAMN w/o ECAF" refers to removing extensible co-attention fusion method and directly concatenating multi-level attentions for fusion.

From the comparison results in Tables 1 and 2, combined with the ablation experimental results in Table 3, we have the following observations:

1) By utilizing multi-level attentions maps to filter noise for multi-granularity features and extensibly fusing them, the proposed MAMN model achieves the best performance on all three datasets. In terms of accuracy, compared with the state-of-the-art methods MMIM and Self-MM, our model achieves the best performance at 76.57%, 78.34% and 75.41% with increases of 3.31%, 3.15% and 11.48% on MVSA-Single, MVSA-Multi and Multi-ZOL datasets, respectively. As for the F1-measure, compared with the state-of-the-art method Self-MM, our MAMN model achieves the best performance at 76.08%, 77.92% and 74.79% with increases of 3.25%, 3.04% and 11.34% on MVSA-Single, MVSA-Multi and Multi-ZOL datasets.

   The reasons that the proposed MAMN outperforms other models on those three datasets are given as follows: (a) it sufficiently extracts multi-granularity features for both images and texts; (b) multi-level attention maps are designed to filter noise and enhance the representation ability for the original multi-granularity features; (c) the extensible co-attention fusion mechanism is utilized to fuse various multi-level attention maps, which can extensibly and effectively interact the correlations contained in any number of multi-granularity features.
2) The performance of "MAMN-Glove" is slightly dropped if Glove is used to extract text features, however, the performance of the MAMN model is

TABLE 3
Ablation Experiment Results

Models	MVSA-Single			MVSA-Multi			Multi-ZOL		
	Acc	Wei-F1	Mac-F1	Acc	Wei-F1	Mac-F1	Acc	Wei-F1	Mac-F1
w/o MG	73.35	72.76	72.71	74.57	74.11	73.96	68.75	68.49	68.46
w/o text	73.96	73.56	74.48	75.82	75.38	75.34	70.65	70.21	70.15
w/o image	74.82	74.55	74.50	76.26	75.93	75.85	72.22	71.86	71.74
MAMN w/o AM	71.85	71.44	71.38	73.61	73.02	73.95	71.18	70.72	70.69
w/o Multi	73.28	72.69	72.65	75.53	75.10	74.97	72.82	72.47	72.41
w/o Single	74.73	74.36	74.29	76.47	75.88	75.80	73.67	73.13	72.08
w/o Main	74.13	73.65	73.57	75.52	75.11	75.04	72.83	72.42	72.36
MAMN w/o ECAF	67.21	66.58	66.51	69.52	68.94	68.89	61.47	61.12	61.04
MAMN	**76.57**	**76.08**	**76.02**	**78.34**	**77.92**	**77.87**	**75.41**	**74.83**	**74.79**

still better than all the baselines in Tables 1 and 2. This fact indicates that the pre-trained language model has a little effect on the performance of the MANMN model.

3) Compared the performance of "MAMN w/o AM" in Table 3 with the baselines, it is observed that when multi-granularity features or multi-level attention maps are removed, the performance of "MAMN w/o AM" model drops about 5%. However, the performance of "MAMN w/o AM" is better than those of most baselines. It is reasonable since original multi-granularity features still contain useful information for fusion, and can be extensibly and efficiently fused by extensible co-attention fusion method.

4) From the performance of "MAMN w/o ECAF" model, the performance of the MAMN model drops a lot after the removal of extensible co-attention fusion method, since multimodal fusion is essential to multimodal sentiment analysis and simple concatenating operations cannot effectively explore the associated information between multiple modalities.

4.3 Ablation Experiments

To verify the effectiveness of every module in our model, eight variants of our MAMN model are designed.

Table 3 shows the results of ablation experiments. Here, "w/o text" model means removing the word-level features of texts. "w/o image" model refers to removing the scene features and object features of images. "w/o MG" model leverages the sentence-level features of texts and the global features of images. Additionally, "w/o Single" model uses the features directly extracted from modality data instead of text attention map and image attention map for multimodal feature fusion. "w/o Multi" model directly concatenates multi-granularity features of texts and images instead of multimodal attention map. Moreover, EMFH pooling mechanism is utilized in both text attention module and image attention module in "w/o Main" model. And "MAMN w/o AM" and "MAMN w/o ECAF" models have the same meanings as in Section 4.2. "Wei-F1" and "Mac-F1" represent weighted-F1 and macro-F1, respectively.

From Table 3, it is seen that "MAMN w/o AM" model performs worse than "w/o Single" and "w/o Multi" models. This fact shows that the introduction of multi-level attention maps can improve the performance for MSA. Moreover, the result of "MAMN w/o ECAF" model is the worst in all the three datasets, which illustrates the importance of extensible co-attention fusion method and EMFH mechanism. In addition, the performance of "w/o MG" model is worse than those of "w/o text" and "w/o image" models, which shows the necessity of multi-granularity features. And the performance of "w/o Main" model is not as good as MAMN model, which demonstrates the effectiveness of the main modality in a specific scenario.

4.4 Hyperparameter Sensitivity Analysis

A series of hyperparameter experiments are conducted to analyze the sensitivity of four hyperparameters, including batch size in {8, 16, 32, 64, 128, 256, 512, 1024}, learning rate in {0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1}, number of cascaded EMFB blocks in {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, and output vector length of the EMFH in {300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200}.

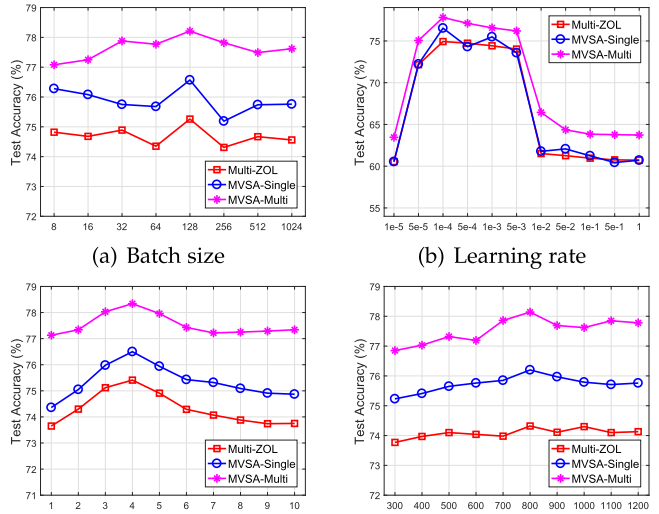


Fig. 4. Hyperparameter sensitivity analysis results.

Fig. 4 shows the results of the hyperparameter sensitivity analysis. It can be seen that the MAMN model has little sensitivity to batch size and output vector length of the fusion modules. The number of cascaded EMFB blocks has a certain impact on the prediction accuracy. Additionally, the model is sensitive to the learning rate.

4.5 Visualization

Actually, an important advantage of the MAMN model lies in that it can explicitly mine the contribution of multi-level attention maps from multi-granularity features for multimodal fusion. Fig. 5 illustrates two examples where darker color represents higher contribution and vice versa.

The following observations can be obtained from Fig. 5:

1) The attention maps from different levels have distinct contributions to multimodal fusion. The multimodal attention map has more contributions than the other single-modality attention maps. In terms of the multi-level attention maps, the text sentence-level attention map in Fig. 5a has the most contribution, and the image scene attention map in Fig. 5b has the largest contribution. This is reasonable since the whole sentence in Fig. 5a contains richer emotional information, while the sentiment information is mainly extracted from the bright and informative scenes of the images in Fig. 5b.

2) The attention features provide more contributions than the corresponding single-modality attention maps. This fact indicates that the interactions among multi-level attention maps have been effectively mined in attention map fusion module. In addition, after the second fusion of EMFH, the fusion feature has the highest contribution, compared with multi-level attention maps and attention features. Because the fusion feature integrates the exclusive and correlated information of multi-level attention maps and attention features. This also indicates the effectiveness of extensible co-attention fusion method.

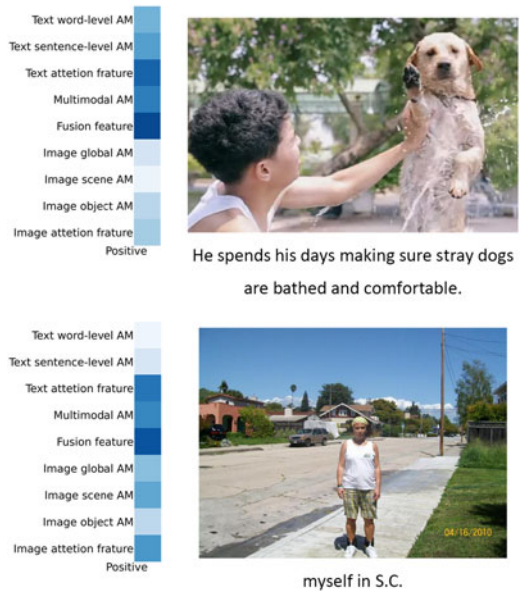


Fig. 5. Visualization of the contribution of multi-level attention maps and attention features for multimodal sentiment analysis.

5 Conclusion

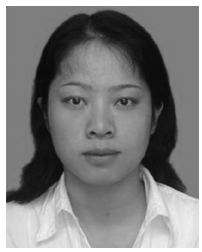
A multi-level attention map network (MAMN) for multimodal sentiment analysis has been proposed to filter noise within and among multi-granularity features, and to fuse the various filtered features in an extensible and efficiently way. Different from previous works, multi-level attention maps based on multi-granularity features are generated before fusing multimodal features, which can filter noise and enhance the representation ability for multi-granularity features. Additionally, the extensible co-attention fusion method is proposed to extensibly and efficiently perform multimodal feature fusion for multi-level attention maps. Moreover, our model shows good extensibility and is able to deal with any number of various multi-granularity features. Extensive experiments have been conducted and the experimental results show that our model has achieved the state-of-the-art performance on three multimodal sentiment analysis datasets. In the future, we plan to design the fusion model with the residual mechanism to solve multimodal tasks such as image captioning.

References

[1] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
[2] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850
[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[4] Y. Zhu, Q. Lin, H. Lu, K. Shi, P. Qiu, and Z. Niu, "Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks," *Knowl. Based Syst.*, vol. 215, 2021, Art. no. 106744.
[5] K. Shi, Y. Wang, H. Lu, Y. Zhu, and Z. Niu, "EKGTF: A knowledge-enhanced model for optimizing social network-based meteorological briefings," *Inf. Process. Manag.*, vol. 58, no. 4, 2021, Art. no. 102564.
[6] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4477–4481.
[7] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 1–14, 2020.
[8] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 10 790–10 797.
[9] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2015, vol. 9362, pp. 159–167.
[10] D. Ghosal, M. S. Akhtar, D. S. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 3454–3466.
[11] D. Gkoumas, Q. Li, Y. Yu, and D. Song, "An entanglement-driven fusion neural network for video sentiment analysis," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1736–1742.
[12] 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. 2, 2016, Art. no. 41.
[13] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 371–378.
[14] N. Xu and W. Mao, "MultiSentiNet: A deep semantic network for multimodal sentiment analysis," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 2399–2402.
[15] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.
[16] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, vol. 10, pp. 2200–2204.
[17] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.
[18] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys," in *Proc. 23rd Int. Conf. Comput. Linguistics, Posters*, 2010, pp. 241–249.
[19] A. Hu and S. R. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 350–358.
[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[21] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[22] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, "Image and encoded text fusion for multi-modal classification," in *Proc. Digit. Image Comput.: Techn. Appl.*, 2018, pp. 1–7.
[23] 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.
[24] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.
[25] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[26] D. Gkoumas, Q. Li, S. Dehdashti, M. Melucci, Y. Yu, and D. Song, "Quantum cognitively motivated decision fusion for video sentiment analysis," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2021, pp. 827–835.
[27] D. S. Chauhan, D. S. R. A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4351–4360.
[28] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[29] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[30] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[32] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1571–1581.
[33] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[34] Y. Li, N. Wang, J. Liu, and X. Hou, "Factorized bilinear models for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2098–2106.
[35] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining*, 2010, pp. 995–1000.
[36] D. Tao, J. Cheng, M. Song, and X. Lin, "Manifold ranking-based matrix factorization for saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1122–1134, Jun. 2016.
[37] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang, "Person re-identification by dual-regularized KISS metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2726–2738, Jun. 2016.
[38] D. Tao, L. Jin, Y. Yuan, and Y. Xue, "Ensemble manifold rank preserving for acceleration-based human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1392–1404, Jun. 2016.
[39] T. Niu, S. Zhu, L. Pang, and A. El-Saddik, "Sentiment analysis on multi-view social data," in *Proc. Int. Conf. Multimedia Model.*, 2016, vol. 9517, pp. 15–27.
[40] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.
[41] K. Zhang, Y. Geng, J. Zhao, J. Liu, and W. Li, "Sentiment analysis of social media via multimodal feature fusion," *Symmetry*, vol. 12, no. 12, 2020, Art. no. 2010.
[42] D. Borth, R. Ji, T. Chen, T. M. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
[43] C. Baecchi, T. Uricchio, M. Bertini, and A. D. Bimbo, "A multimodal feature learning approach for sentiment analysis of social network multimedia," *Multimedia Tools Appl.*, vol. 75, no. 5, pp. 2507–2525, 2016.
[44] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, 2017, pp. 152–154.
[45] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 1–7.
[46] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 9180–9192.
[47] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 606–615.
[48] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 214–224.
[49] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 4068–4074.
[50] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 452–461.
[51] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

Xiaojun Xue received the BE degree from the China University of Geosciences, Beijing, China, in 2019. He is currently working toward the PhD degree in the School of Computer Science and Technology, Beijing Institute of Technology, China. His research interests include social network analysis, data mining, information extraction, and machine learning.



Chunxia Zhang received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2005. As an academic visitor, she visited the University of Vermont, Burlington, Vermont from January 2010 to February 2011. She is currently an associate professor with the School of Computer Science and Technology, Beijing Institute of Technology, China. Her research interests include information extraction, social computing, and machine learning and so on.

Zhendong Niu received the PhD degree in computer science from the Beijing Institute of Technology, Beijing, China, in 1995. From 1996 to 1998, he was a postdoctoral researcher with the University of Pittsburgh, Pittsburgh, Pennsylvania, where he has been a joint professor with the School of Computing and Information since 2006. He was a researcher and adjunct faculty member with Carnegie Mellon University, Pittsburgh, Pennsylvania, from 1999 to 2004. He is currently a professor with the School of Computer Science and Technology, Beijing Institute of Technology, China. His current research interests include informational retrieval, software architecture, digital libraries, and Web-based learning techniques. He was a recipient of the IBM Faculty Innovation Award, in 2005 and the New Century Excellent Talents in the University of Ministry of Education of China, in 2006.

Xindong Wu (Fellow, IEEE) received the BS and MS degrees in computer science from the Hefei University of Technology, Hefei, China, and the PhD degree in artificial intelligence from the University of Edinburgh, Edinburgh, Britain. He is a Yangtze River scholar with the School of Computer Science and Information Engineering, Hefei University of Technology, China, and the president of the Mininglamp Academy of Sciences, Minininglamp, Beijing, China, and a fellow of the AAAS. His research interests include data mining, big data analytics, knowledge-based systems, and Web information exploration. He is currently the steering committee chair of the IEEE International Conference on Data Mining (ICDM), the editor-in-chief of the *Knowledge and Information Systems (KAIS, by Springer)*, and a series editor-in-chief of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He was the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering* (TKDE, by the IEEE Computer Society) between 2005 and 2008. He served as program committee chair/co-chair for the 2003 IEEE International Conference on Data Mining, the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, and the 19th ACM Conference on Information and Knowledge Management.