

# Cross-Modal Complementary Network with Hierarchical Fusion for Multimodal Sentiment Classification

Cheng Peng, Chunxia Zhang\*, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu

**Abstract:** Multimodal Sentiment Classification (MSC) uses multimodal data, such as images and texts, to identify the users' sentiment polarities from the information posted by users on the Internet. MSC has attracted considerable attention because of its wide applications in social computing and opinion mining. However, improper correlation strategies can cause erroneous fusion as the texts and the images that are unrelated to each other may integrate. Moreover, simply concatenating them modal by modal, even with true correlation, cannot fully capture the features within and between modals. To solve these problems, this paper proposes a Cross-Modal Complementary Network (CMCN) with hierarchical fusion for MSC. The CMCN is designed as a hierarchical structure with three key modules, namely, the feature extraction module to extract features from texts and images, the feature attention module to learn both text and image attention features generated by an image-text correlation generator, and the cross-modal hierarchical fusion module to fuse features within and between modals. Such a CMCN provides a hierarchical fusion framework that can fully integrate different modal features and helps reduce the risk of integrating unrelated modal features. Extensive experimental results on three public datasets show that the proposed approach significantly outperforms the state-of-the-art methods.

**Key words:** multimodal sentiment analysis; multimodal fusion; Cross-Modal Complementary Network (CMCN); hierarchical fusion; joint optimization

## 1 Introduction

In the current age, more users post emotional content and opinions on social platforms, such as Weibo and Twitter. Sentiment classification with multimodal data (i.e., texts, images, and/or videos) plays an important role in the tasks of public opinion monitoring, advertisement

recommendation, and hot event detection.

Multimodal Sentiment Classification (MSC) is a fundamental issue in the fields of social computing and sentiment analysis. The MSC approaches can be roughly divided into two kinds, namely, methods with interactions between multiple modalities<sup>[1–3]</sup> and methods without interactions when extracting feature vectors<sup>[4–6]</sup>. For instance, Xu et al.<sup>[1]</sup> utilized two interactive networks to learn both interactive influences between cross-modality data and the influences themselves in single-modality data simultaneously. Moreover, Yu et al.<sup>[2, 3]</sup> designed the multimodal factorized bilinear pooling approach to fuse multimodal features and constructed the co-attention mechanism to learn the text and image attentions. In parallel, Hazarika et al.<sup>[4]</sup> and Hu and Flaxman<sup>[5]</sup> directly concatenated multiple features, including text, image, and audio

• Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, and Zhengdong Niu are with School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail: 3120191033@bit.edu.cn; cxzhang@bit.edu.cn; 3120205504@bit.edu.cn; 3220190942@bit.edu.cn; zniu@bit.edu.cn.

• Hongjian Liang is with School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan. E-mail: yuanhenglj@akane.waseda.jp.

\* To whom correspondence should be addressed.

Manuscript received: 2021-05-14; revised: 2021-07-15; accepted: 2021-07-30

features, and performed sentiment classification on the concatenated features. Furthermore, Anderson et al.<sup>[6]</sup> transformed the text and image features into vectors with the same dimensions and dealt with feature fusion with the dot product.

However, some recent works on MSC utilize erroneous correlations between different modalities and relatively simple feature fusion methods, such as dot product and concatenation<sup>[7, 8]</sup>. For instance, after Obama won the 2012 presidential election, he posted the textual content “Four more years” on his official Twitter. The auxiliary image is a big hug of him and his wife. Within this tweet, people can understand that the president’s successful election for four more years is related to the happy embrace. However, the features of the period “four years” and the features of a hug are not necessarily related to each other. For such a pair of samples, the degree of fusion between modalities with cross-modal attention needs to be reduced. Moreover, previous approaches seldom consider utilizing the interaction information within a single modality and between modalities at the same time. Actually, different modalities can benefit from each other when capturing implicit semantic features because complementary relationships exist between different modalities<sup>[7]</sup>.

To solve the aforementioned problems, this paper proposes a Cross-Modal Complementary Network (CMCN) with hierarchical fusion for MSC. The main contributions can be summarized as follows:

(1) A CMCN with hierarchical fusion is developed for the MSC task. The CMCN is designed as a hierarchical structure with three key modules, namely, the Feature Extraction Module (FEM), Feature Attention Module (FAM), and Cross-Modal Hierarchical Fusion module (CMHF). With the help of an image-text correlation generator, the CMCN can avoid the fusion of irrelevant images and texts mutual-modal correlations.

(2) The image-text correlation generator is constructed to measure the similarity between images and texts for learning image attention features. To this end, a Visual Geometry Group Network (VGGNet) is employed to learn image features, whereas a Bidirectional Encoder Representation from Transformers (BERT) is used to learn text features. Moreover, an image caption generation strategy is utilized to convert the images into descriptive texts. Furthermore, the image-text correlation generator computes the correlation degree between original texts and descriptive texts of images

as image-text correlations. Typically, the image-text correlations are fused with the text attention features and the image features to generate the image attention features, which can reduce the influences caused by erroneous image-text correlations.

(3) A CMHF is proposed to fuse four kinds of features (i.e., text features, image features, text attention features, and image attention features) within a single modality and between multiple modalities hierarchically. This module consists of four layers, namely, feature sampling, cross-modal fusion, global fusion, and classification layers. The feature sampling layer upsamples those four kinds of features to the same dimension. The cross-modal fusion layer implements fusion between original features of images and texts, between attention features of images and texts, and between original and attention features within a single modality. Furthermore, the global fusion layer integrates all of the fused features outputted by the cross-modal fusion layer. The loss of each layer is used for joint optimization of the global model. The CMHF provides a hierarchical mechanism to fuse those four types of features with the joint optimization method, which can capture the implicit features within a single modality and the complementary features between multiple modalities.

(4) Extensive experiments are conducted on three public datasets, namely, MVSA-single, MVSA-multiple<sup>[9]</sup>, and Multi-ZOL<sup>[11]</sup> datasets. The experimental results show that the performance of our method is superior to those of other related works, and the ablation experiments indicate that each component of our model contributes to the sentiment classification results.

The rest of this paper is organized as follows. Section 2 introduces the related works of MSC. Section 3 presents the proposed CMCN model. Section 4 gives the experimental results. Section 5 concludes this paper.

## 2 Related Work

Preliminary works on the MSC task extracted suitable features for sentiment classification<sup>[7, 8]</sup>. Recently, most works employ Convolutional Neural Networks (CNN) or their variants to extract image features and use the output of the hidden layers of those networks as image features for subsequent neural network input<sup>[10–12]</sup>. The majority of works utilized BERT<sup>[13]</sup> for text feature extraction and Long Short-Term Memory (LSTM)<sup>[5, 6]</sup> for feature extraction. BERT can learn distributed feature representation for words by running a self-supervised learning method on a large-scale corpus. Furthermore, some works extracted the features of a

single modality with the aid of other modal data. For example, Xu and Mao<sup>[14]</sup> built the MultiSentiNet neural network, which uses visual features to guide the attention LSTM model to extract text features, and classified the concatenated vectors of visual and text features. Then, Xu et al.<sup>[15]</sup> proposed a co-memory network to achieve interactive sentiment classification between texts and images iteratively.

In the early stage, text-based methods are usually designed for sentiment classification. These methods can be classified into three kinds<sup>[7]</sup>, namely, document-level, sentence-level, and aspect-level sentiment classification. The three related works on these three kinds of methods are as follows: Xu et al.<sup>[16]</sup> obtained the overall semantic information for the entire document using the cached LSTM model. Then, they used the memory groups with low forgetting rates to build the global features and the memory groups with high forgetting rates to obtain the local features. Moreover, Mishra et al.<sup>[17]</sup> learned the eye movement features from human reading behavior by utilizing the CNN and analyzed the sentiment from the eye movement and textual features. Ma et al.<sup>[18]</sup> proposed an interactive attention network for aspect-level sentiment classification. They used two different LSTMs to learn the important words from the target and full texts interactively and concatenate them for sentiment classification. Furthermore, Zhao et al.<sup>[8]</sup> introduced other unimodal sentiment analysis methods, such as AlexNet, for sentiment analysis of images and mid-level perceptual musical feature modeling for sentiment analysis of music.

Recently, many users publish their daily content on the Internet in various forms, such as texts and images. Thus, the existing MSC works mainly focus on two kinds of datasets, namely, text-image and video datasets. For the text-image datasets, Gaspar and Alexandre<sup>[19]</sup> segmented the content that was posted on social media into texts, images, and image contents for sentiment classification to reduce the dependency of text analysis. Truong and Lauw<sup>[20]</sup> proposed VistaNet to analyze textual and visual components. The texts were divided into multiple granularities, i.e., word, sentence, and document levels, and the images were divided into multiple aspects, such as objects or entities. Moreover, Liu et al.<sup>[21]</sup> proposed CASA for performing context-aware user sentiment analysis, which involves the semantic correlation and the effects of context information. For the video datasets, many works initially extracted the images, audios, and texts from the videos<sup>[4, 22–26]</sup>. Furthermore, Majumder

et al.<sup>[25]</sup> utilized hierarchical fusion to fully integrate features of images, audios, and texts. To introduce the interpersonal influences, Hazarika et al.<sup>[4]</sup> proposed an interactive conversational memory network that can model the emotional impact of self-speaker and inter-speaker to obtain global memories. Then, the contextual summaries generated by the global memories helped sentiment prediction.

The performance of sentiment analysis on multiple modalities (i.e., texts, images, videos, and audios) was better than that on a single modality as multimodal data provide more information than single-modality data, such as visual or auditory information<sup>[7]</sup>. The key to MSC is how to design the feature fusion method within a single modality and between multiple modalities. The fusion strategies of multiple modalities can be divided into two types of methods, namely, ones with the interaction between multiple modalities and ones without the interaction between multiple modalities. Yu et al.<sup>[2, 3]</sup> first introduced the multimodal factorized bilinear pooling approach to integrating image and text features and then used the co-attention mechanism to jointly learn the image and text attentions. Moreover, Zhang et al.<sup>[27]</sup> developed semi-supervised variational autoencoders to extract independent knowledge from single-modality data and interactive knowledge from different modalities. Furthermore, Wang et al.<sup>[28]</sup> proposed an end-to-end fusion method to perform the MSC task. The method employed a transformer for modal conversion and fusion, and the fused features contained the source and target modality information at the same time.

The works without the interaction between two modalities include the methods of Cambria et al.<sup>[26]</sup> and Gaspar and Alexandre<sup>[19]</sup>. Gaspar and Alexandre<sup>[19]</sup> performed a weighted summation of the classification results of two modalities. Cambria et al.<sup>[26]</sup> directly concatenated the extracted text, image, and audio features into the fusion features. Yang et al.<sup>[29]</sup> developed common space learning, which can decrease fusion and disagreement problems, to analyze the sentiment tags of micro-videos. The features extracted from visual, acoustic, and textual data are mapped into a common space by MultiLayer Perceptron (MLP) to obtain multimodal features. The Attention-based Modal Gated Network (AMGN) was proposed by Huang et al.<sup>[30]</sup> to classify sentiments. To fuse text and image information, the AMGN utilized modality-gated LSTM to adaptively choose modal features with strong sentiments.

### 3 Method

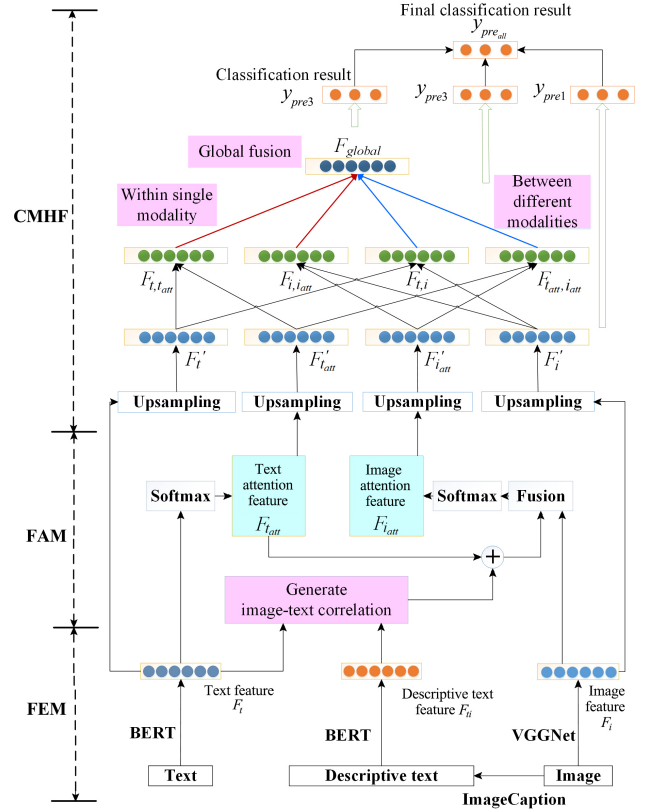
#### 3.1 Problem formulation

The definition of MSC is expressed as follows: the dataset is  $D = \{(s_1, I_1), (s_2, I_2), \dots, (s_n, I_n)\}$ . Here  $I_j = \{i_{j,1}, i_{j,2}, \dots, i_{j,m}\}$  is the image set,  $i_{j,m}$  is the  $m$ -th image of the  $j$ -th tweet or review. Let  $T = \{s_1, s_2, \dots, s_n\}$  be the set of text data of tweets or reviews, where  $s_j$  denotes a tweet or review including only textual data, and  $I = I_1 \cup I_2 \cup \dots \cup I_n$  is the set of images. Here,  $n$  denotes the number of text-image pairs in the datasets. When a tweet or review does not contain images, we use zero vectors to replace image features.

This paper aims to design a function  $f = G(T, I)$ , which can be used to classify the sentiment of multimodal data based on the set  $T$  of texts and the set  $I$  of images. Here  $f$  is the sentiment classification result that we predict and  $G(\cdot)$  is a function from the multimodal data to the sentiment classification result. This paper learns the text features  $F_t$ , image features  $F_i$ , text attention features  $F_{tatt}$ , and image attention features  $F_{iatt}$  to classify the sentiments of multimodal data and perform joint optimization, such that the loss of sentiment prediction is minimized.

#### 3.2 Technical overview

Figure 1 shows the overall architecture of the CMCN with hierarchical fusion for MSC. The CMCN is a hierarchical structure with three key modules, namely, FEM, FAM, and CMHF. (1) The FEM extracts the original features from texts and images and the descriptive textual features from images. (2) The FAM learns both text and image attention features from texts and images. The soft attention mechanism is utilized to learn text attention features. Specifically, the image attention features are semantically generated by the image features, text attention features, and image-text correlation generator. Here, the image-text correlation generator is used to measure the image-text correlations based on the cosine similarity. (3) The CMHF aims to fuse the text features, image features, text attention features, and image attention features within a single modality and between multiple modalities hierarchically. The CMHF consists of four layers, namely, feature sampling, cross-modal fusion, global fusion, and classification layers. The features in each of the first three layers can be concatenated into the vector as the feature of this layer. Finally, the classification layer performs sentiment classification based on the feature vectors of the first three layers



**Fig. 1 Overall architecture of the CMCN with hierarchical fusion for multimodal sentiment classification.**

in the CMHF to obtain three classification results, i.e.,  $y_{pre1}$ ,  $y_{pre2}$ , and  $y_{pre3}$ . Furthermore, the concatenation of the three classification results  $y_{pre1}$ ,  $y_{pre2}$ , and  $y_{pre3}$  is utilized for classification to obtain the final result  $y_{pre-all}$  by softmax, and the losses of the four classification results  $y_{pre1}$ ,  $y_{pre2}$ ,  $y_{pre3}$ , and  $y_{pre-all}$  are used for joint optimization.

#### 3.3 Feature extraction module

In the FEM, the CMCN model generates the original text feature vector  $F_t$  from the texts, original image feature vector  $F_i$  from the images, and descriptive text feature vector  $F_{ti}$  from the descriptive texts of images.

First, the pretraining model BERT<sup>[13]</sup> is used to extract the text feature vector  $F_t$  from the original text  $T$ ,

$$F_t = BERT(T) \quad (1)$$

Second, we utilize VGGNet<sup>[10]</sup> to build the image feature vector  $F_i$  from image  $I$ ,

$$F_i = VGGNet(I) \quad (2)$$

Third, the image caption generation method which is designed by Xu et al.<sup>[31]</sup> is employed to convert image  $I$  into the descriptive text  $T_i$  of the images. This method can automatically learn how to describe the contents of the images, visually show how the attention is

distributed on the objects within the images, and generate corresponding words at the same time. Then, the CMCN model uses BERT to build the descriptive text feature vector  $F_{ti}$  from the descriptive text  $T_i$ ,

$$T_i = \text{ImageCaption}(I) \quad (3)$$

$$F_{ti} = \text{BERT}(T_i) \quad (4)$$

### 3.4 Feature attention module

In contrast to images formatted with pixels that are only related to low-level physical features, text information contains high-level semantic features, and text attention features are more discriminative and semantic for sentiment classification. Thus, the texts are considered as the main modality and are employed to help guide the learning of the attention features of the images. The FAM first builds the attention feature vectors of the texts, and then generates the image attention feature vectors using the image features, text attention features, and image-text correlation generator.

#### 3.4.1 Text attention feature learning

In the sentences, not all words contribute equally to sentiment classification, and some words have more distinct features for sentiment detection. Therefore, different weights need to be assigned to each word when inferring the sentiment polarity of the texts. A soft attention mechanism was introduced by Xu et al.<sup>[1]</sup> to learn the weights of words,

$$h_k = \text{ReLU}(\mathbf{W}_k t_k + b_k), t_k \in F_t \quad (5)$$

$$\alpha_k = \frac{\exp(h_k)}{\sum_p \exp(h_p)} \quad (6)$$

$$F_{tatt} = \sum_k \alpha_k t_k \quad (7)$$

In Eq. (5), a neuron layer with a nonlinear activation function ReLU is used to project each word embedding  $t_k$  in  $h_k$  in the sentence.  $\mathbf{W}_k$  and  $b_k$  are the weight matrix and bias vector, respectively. Then, the softmax is utilized to generate the attention weight  $\alpha_k$  in Eq. (6).

Finally, the word embedding  $t_k$  of each word in the sentence is weighted and summed according to its attention weight  $\alpha_k$  to obtain the text attention feature vector  $F_{tatt}$  in Eq. (7).

#### 3.4.2 Image attention feature learning

The CMCN model introduces the image-text correlation generator to learn the importance of text attention feature vectors in the generation process of image attention feature vectors. Notably, the erroneous correlations

between different modalities may cause the fusion of unrelated images and texts. Hence, the image-text correlation is introduced to measure the influence of text attention feature vectors in the generation process of image attention feature vectors. The text attention feature vector  $F_{tatt}$ , image feature vector  $F_i$ , and their correlation  $c$  are fused to generate the image transition vector  $F_{itrans}$ ,

$$F_{itrans} = F_i + \theta \times \sigma(F_{tatt}, c) \quad (8)$$

$$c = \cos(F_t, F_{ti}) \quad (9)$$

$$\sigma(F_{tatt}, c) = F_{tatt} \times (\text{norm}(c) + 0.5) \quad (10)$$

In Eq. (8),  $\theta$  is a hyperparameter, which determines the degree of influence of the text attention feature vectors on the image feature vectors. When the text and image information is equally important, the  $\theta$  value is set as 1, and the larger  $\theta$ , the more important the text information. The correlation  $c$  in Eq. (9) is obtained based on the cosine similarity between the original text feature vector  $F_t$  and the descriptive text feature vector  $F_{ti}$ . The reason that we choose to use two kinds of text features ( $F_t$ ,  $F_{ti}$ ) to calculate the correlation  $c$  in Eq.(9) is that the correlation calculated from the features within the same modal with the same model is more reliable. Moreover, the correlation calculated by  $F_t$  and  $F_{ti}$  has better experimental results than the correlation calculated by the text feature  $F_t$  learned by BERT and image feature  $F_i$  learned by VGGNet. In Eq. (10),  $\text{norm}(\cdot)$  is the zero-mean normalization.

The soft attention mechanism<sup>[1]</sup> is also employed to generate the image attention feature vector  $F_{iatt}$  based on

$$h_l = \text{ReLU}(\mathbf{W}_l t_l + b_l), t_l \in F_{itrans} \quad (11)$$

$$\alpha_l = \frac{\exp(h_l)}{\sum_q \exp(h_q)} \quad (12)$$

$$F_{iatt} = \sum_l \alpha_l \times t_l \quad (13)$$

### 3.5 CMHF module

To fuse the features within a single modality and between multiple modalities, the CMHF is composed of four layers, namely, feature sampling, cross-modal fusion, global fusion, and classification layers.

The feature sampling layer upsamples the four kinds of feature vectors  $F_t$ ,  $F_i$ ,  $F_{tatt}$ , and  $F_{iatt}$  to the same dimension to ensure that these feature vectors can be integrated more accurately. The formulas are expressed

as follows:

$$F'_t = \text{Relu}(\mathbf{W}_t F_t + b_t) \quad (14)$$

$$F'_i = \text{Relu}(\mathbf{W}_i F_i + b_i) \quad (15)$$

$$F'_{tatt} = \text{Relu}(\mathbf{W}_{tatt} F_{tatt} + b_{tatt}) \quad (16)$$

$$F'_{iatt} = \text{Relu}(\mathbf{W}_{iatt} F_{iatt} + b_{iatt}) \quad (17)$$

The cross-modal fusion layer conducts four fusion operations, namely,  $g(F'_t, F'_i)$ ,  $g(F'_{tatt}, F'_{iatt})$ ,  $g(F'_i, F'_{iatt})$ , and  $g(F'_{tatt}, F'_{iatt})$ , as expressed in Eqs. (18)–(21), respectively,

$$F_{t,i} = g(F'_t, F'_i) \quad (18)$$

$$F_{tatt,iatt} = g(F'_{tatt}, F'_{iatt}) \quad (19)$$

$$F_{t,tatt} = g(F'_t, F'_{tatt}) \quad (20)$$

$$F_{i,iatt} = g(F'_i, F'_{iatt}) \quad (21)$$

where  $F_{t,i}$  is obtained by the fusion of  $F'_t$  and  $F'_i$  using the way of dot product, whereas  $F_{tatt,iatt}$  is constructed by the fusion of  $F'_{tatt}$  and  $F'_{iatt}$ . Typically,  $F_{t,tatt}$  can be obtained through the integration operation between  $F'_t$  and  $F'_{tatt}$ , whereas  $F_{i,iatt}$  is acquired by fusing  $F'_i$  and  $F'_{iatt}$ .

The global fusion layer is used to fuse all of the fused feature vectors outputted by the cross-modal fusion layer,

$$F_{global} = g(F_{t,i}, F_{tatt,iatt}, F_{t,tatt}, F_{i,iatt}) \quad (22)$$

where  $g(F_{t,i}, F_{tatt,iatt}, F_{t,tatt}, F_{i,iatt})$  means that  $F_{t,i}$ ,  $F_{tatt,iatt}$ ,  $F_{t,tatt}$ , and  $F_{i,iatt}$  are integrated by the dot product.

Finally, the CMCN model concatenates the feature vectors of each of the first three layers of CMHF and utilizes softmax for classification. Thus, we obtain three classification results, i.e.,  $y_{pre1}$ ,  $y_{pre2}$ , and  $y_{pre3}$ . Then, the CMCN model concatenates and classifies  $y_{pre1}$ ,  $y_{pre2}$ ,  $y_{pre3}$  to obtain  $y_{preall}$ ,

$$y_{pre1} = \text{softmax}(\text{Relu}(\mathbf{W}_1[F'_t, F'_i, F'_{tatt}, F'_{iatt}] + b_1)) \quad (23)$$

$$y_{pre2} = \text{softmax}(\text{Relu}(\mathbf{W}_2[(F_{t,i}, F_{tatt,iatt}, F_{t,tatt}, F_{i,iatt}] + b_2)) \quad (24)$$

$$y_{pre3} = \text{softmax}(\text{Relu}(\mathbf{W}_3 F_{global} + b_3)) \quad (25)$$

$$y_{preall} = \text{softmax}(\text{Relu}(\mathbf{W}_4[y_{pre1}, y_{pre2}, y_{pre3}] + b_4)) \quad (26)$$

Then, the losses of the classification results (i.e.,  $(y_{pre1}, y_{pre2}, y_{pre3}, \text{ and } y_{preall})$ ) and the expected results  $y$  are calculated using the cross-entropy error function,

$$L(F_i) = \text{CrossEntropyLoss}(y_{prei}, y), \quad i \in [1, 2, 3] \quad (27)$$

$$L(F_4) = \text{CrossEntropyLoss}(y_{preall}, y) \quad (28)$$

Furthermore, the joint optimization method is employed to optimize the losses of the classification results. Through the joint optimization method, the proposed model can train the features of different granularities of the first three layers of CMHF and obtain better experimental results,

$$L = L(F_1) + L(F_2) + L(F_3) + L(F_4) + \lambda \|\theta\|_2 \quad (29)$$

The proposed model is trained by minimizing the total loss  $L$  with the Adam optimization algorithm. The regularization parameter  $\lambda$  is set as 0.0001. Finally, our model uses  $y_{preall}$  as the classification result for the calculation of accuracy and the F1 measure.

## 4 Experimental Setting

### 4.1 Datasets

The CMCN model in this paper is evaluated on three public datasets, namely, MVSA-single, MVSA-multiple<sup>‡</sup>, and Multi-ZOL<sup>§</sup> datasets. The MVSA-single and MVSA-multiple datasets were used in the work of Niu et al.<sup>[9]</sup>, and the Multi-ZOL dataset was constructed by Xu et al.<sup>[11]</sup>. Table 1 shows the details of the three datasets, and Tables 2 and 3 shows the distribution of sentiment polarity labels in each dataset. The experimental task of this paper is to classify the

**Table 1 Overall statistics of the datasets.**

Dataset	Number of original text-image pairs	Number of processed text-image pairs	Number of sentiment polarities
MVSA-single	5129	4511	3
MVSA-multiple	19600	17024	3
Multi-ZOL	5288	28469	10

**Table 2 Number of text-image pairs for each sentiment polarity in the MVSA-single and MVSA-multiple datasets.**

Sentiment polarity	Number of text-image pairs	
	MVSA-single	MVSA-multiple
Positive	2683	11318
Negative	1358	1298
Neutral	470	4408
Total	4511	17024

<sup>‡</sup><http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/>.

<sup>§</sup><https://github.com/xunan0812/MIMN>.

**Table 3** Number of text-image pairs for each sentiment polarity in the Multi-ZOL dataset.

Sentiment polarity	Number of text-image pairs
1	6
2	1434
3	4
4	1491
5	3
6	4355
7	0
8	8695
9	0
10	12 481
Total	28 469

sentiment polarity of multimodal data. Specifically, the number of sentiment polarity classification labels of the three datasets are 3, 3, and 10, as shown in Table 1.

The MVSA-single dataset is composed of 5129 text-image pairs on Twitter. Each text-image pair has text and image labels that were annotated by an annotator<sup>[9]</sup>. When the sentiment polarity of the text label is opposite to that of the image label, the text-image pair is eliminated. Finally, the MVSA-single dataset is left with 4511 text-image pairs.

The MVSA-multiple dataset consists of 19 600 text-image pairs, and each text-image pair is annotated by three annotators<sup>[8]</sup>. Finally, after eliminating the text-image pairs with opposite text and image labels, the MVSA-multiple dataset is left with 17 024 text-image pairs.

The Multi-ZOL dataset is composed of popular mobile phone reviews crawled from ZOL.com<sup>[1]</sup>, with a total of 5288 multimodal reviews. Each multimodal comment comprises one text content, some images, and one to six aspects. Each aspect of the review can be scored 1–10 points. According to each aspect, each multimodal review is divided into a set of triples (i.e., aspect, the corresponding text, and the corresponding images). Finally, we obtain a total of 28 469 triples. In this paper, the three datasets are randomly divided into 80% training set, 10% validation set, and 10% test set.

All experiments are conducted under the framework of the development environment PyTorch 1.3.1.

## 4.2 Baseline methods

To demonstrate the effectiveness of the proposed CMCN model, we compare the performance of the CMCN model with those of the multimodal sentiment analysis baseline and the state-of-the-art methods in Tables 4 and 5. On the MSVA-single and MVSA-multiple

**Table 4** Comparison of methods on the MVSA dataset. (%)

Method	MVSA-single		MVSA-multiple	
	Accuracy	Weighted-F1	Accuracy	Weighted-F1
SentiBank+				
SentiStrength				
CBOW+DA+				
LR				
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
MN-Hop2+				
img2text	68.07	65.19	67.92	67.16
CoMN-Hop4	69.18	68.29	69.92	69.83
CoMN-Hop6	70.51	70.01	68.92	68.83
CFE-ATT	71.44	71.06	69.62	69.35
MVAN-M	72.98	72.98	<b>72.36</b>	72.30
Concatenation	64.30	65.78	66.22	67.50
GMU	64.52	69.52	65.92	73.45
MFB	71.62	72.98	69.68	72.56
CMCN	<b>73.61</b>	<b>75.03</b>	70.45	<b>74.77</b>

**Table 5** Comparison of methods on the Multi-ZOL dataset. (%)

Method	Accuracy	Macro-F1
ATAE-LSTM	59.58	58.95
MemNet	59.51	58.73
IAN	60.08	59.47
RAM	60.18	59.68
Co-MN + Aspect	60.43	59.74
MIMN	61.59	60.51
Concatenation	61.42	47.02
GMU	61.64	42.87
MFB	60.75	37.56
CMCN	<b>74.28</b>	<b>71.51</b>

datasets, we compare our model with SentiBank + SentiStrength<sup>[32]</sup>, CBOW + DA + LR<sup>[33]</sup>, CNN-Multi<sup>[34]</sup>, DNN-LR<sup>[35]</sup>, HSAN<sup>[36]</sup>, MultiSentiNet<sup>[14]</sup>, MN-Hop2 + img2text<sup>[15]</sup>, CoMN-Hop4<sup>[15]</sup>, CoMN-Hop6<sup>[15]</sup>, CFE-ATT<sup>[37]</sup>, MVAN-M<sup>[38]</sup>, Concatenation<sup>[39]</sup>, GMU<sup>[40]</sup>, and MFB<sup>[2]</sup>, as shown in Table 4.

- SentiBank + SentiStrength<sup>[32]</sup> uses SentiBank to extract multiple adjective-noun pairs from images and SentiStrength to determine the sentiment polarity of the text.

- CBOW + DA + LR<sup>[33]</sup> obtains visual information from a large-scale corpus through unsupervised learning and combines that information with the language model based on neural networks for multimodal sentiment analysis.

- CNN-Multi<sup>[34]</sup> utilizes two CNN models to extract

text and image features and uses those features to analyze the sentiment.

- DNN-LR<sup>[35]</sup> employs two CNN models to extract text and image features and utilizes those features to analyze the sentiment by logistic regression.

- HSAN<sup>[36]</sup> obtains the captions of the images and adopts the attention network to deal with the texts and captions together.

- MultiSentiNet<sup>[14]</sup> extracts visual information, such as the scene and object features, of the images and uses this visual information to guide the extraction of text features.

- MN-Hop<sup>[15]</sup> adopts the attention mechanism to interactively model the relationship between text and visual memories. MN-Hop2 + img2text is a variant of MN-Hop. Based on MN-Hop, a co-memory network (such as CoMN-Hop4 and CoMN-Hop6) is proposed to iterate the interaction between texts and images to learn the mutual influence.

- CFF-ATT<sup>[37]</sup> removes the influence of noise in texts using a denoizing autoencoder, extracts image features using an attention-based variational autoencoder, and learns the internal features of texts and images from each other symmetrically.

- MVAN-M<sup>[38]</sup> uses the constantly updated memory network to obtain the deep semantic information of texts and images.

- Concatenation<sup>[39]</sup> simply concatenates the features of two modalities and uses MLP for sentiment classification, and its effectiveness has been demonstrated in a variety of applications<sup>[5, 39, 40]</sup>.

- GMU<sup>[40]</sup> learns how modalities affect the activation of the unit by multiplicative gates and finds an intermediate representation of the data of different modalities.

- MFB<sup>[2]</sup> uses a multimodal factorized bilinear pooling approach to combine multimodal features and a co-attention mechanism to jointly learn both the image and question attentions.

On the Multi-ZOL dataset, we compare our model with ATAE-LSTM<sup>[41]</sup>, MemNet<sup>[42]</sup>, IAN<sup>[18]</sup>, RAM<sup>[43]</sup>, Co-MN + Aspect<sup>[15]</sup>, MIMN<sup>[1]</sup>, Concatenation<sup>[39]</sup>, GMU<sup>[40]</sup>, and MFB<sup>[2]</sup>, as shown in Table 5.

- ATAE-LSTM<sup>[41]</sup> uses LSTM to extract text context information, adds aspect embedding to each word, and fuses the word embedding and aspect embedding in the attention layer to learn the sentiment polarity.

- MemNet<sup>[42]</sup> utilizes the aspect embedding for query and uses a multiple attention mechanism to stack the

word embedding memory into deep memory. The sentiment polarity is classified based on the output of the attention layer.

- IAN<sup>[18]</sup> adopts two different LSTMs to learn the aspect and text attention features and concatenates these features for sentiment classification.

- RAM<sup>[43]</sup> establishes the memory and generates the aspect embedding based on Bi-LSTM. The multiple attention layer is combined with RNN nonlinearly to enhance the expressive capability.

- Co-MN + Aspect<sup>[15]</sup> adds the average value of the aspect embedding to the input of the original text and the image network.

- MIMN<sup>[1]</sup> contains two interactive memory networks, which calculate the texts and images, and learns the mutual influence of cross-modal data and the self-influence of single-modality data at the same time.

### 4.3 Experimental results

In our experiments, the performance of the CMCN model is compared with that of other state-of-the-art models using accuracy, weighted-F1 measure, and macro-F1 measure.

Table 4 shows the experimental results of our CMCN model and the baseline and state-of-the-art methods on the MVSA-single and MVSA-multiple datasets. On the MVSA-single dataset, the CMCN model achieves an accuracy of 73.61%, weighted-F1 of 75.03%, and macro-F1 of 60.51%. On the MVSA-multiple dataset, the CMCN model obtains the accuracy of 70.45%, weighted-F1 of 74.77%, and macro-F1 of 50.58%. Table 4 shows that the CMCN model achieves the best sentiment classification performances on the MVSA-single and MVSA-multiple datasets.

Therefore, we have the following assertions: (1) the baseline SentimentBank + SentiStrength has the worst performance. This finding shows that the usage of deep neural networks can effectively improve the performance of the models. (2) CNN-Multi and DNN-LR use CNN for feature extraction and classification, and their experimental performances are better than that of CBOW + DA + LR. (3) Moreover, HSAN, MultiSentiNet, MN-Hop2 + img2text, CoMN-Hop4, CoMN-Hop6, CFF-ATT, and MVAN-M all introduce the attention mechanism to extract features, and their experimental performances are better than those of the first four methods in Table 4.

For the Multi-ZOL dataset, Table 5 shows the classification performances of other methods and the



CMCN model. The CMCN model achieves the accuracy of 74.28%, weighted-F1 of 74.53%, and macro-F1 of 71.51%. Table 5 also shows that the accuracy of the CMCN model is higher than that of the state-of-the-art method MIMN by 12.69%. Moreover, the macro-F1 of the CMCN model is higher than that of the MIMN by 11%.

Accordingly, we draw the following conclusions: (1) the AEAT-LSTM, MemNet, IAN, and RAM models only analyze text and aspect data; thus, the experimental results of these models are not as good as those of the Co-MN + Aspect, MIMN, and CMCN models that use texts, images, and aspects for analysis. (2) The performance of the CMCN model is significantly better than those of other models because the CMCN model introduces the correlation measure between multiple modalities and a hierarchical fusion mechanism to fully integrate multimodal features within a single modality and between modalities.

#### 4.4 Ablation experiments

##### 4.4.1 Parameter experiments

Figure 2 illustrates the performance of the CMCN model under different parameter values to analyze parameter sensitivity. Experiments are conducted to analyze the hyperparameter sensitivity of our CMCN model, including dropout, learning rate, hidden size, and fusion rate. In the experimental setup for the MVSA-single dataset, the model parameters dropout, learning rate, hidden size, and fusion rate are set as (0.5, 0.0003, 256, 1.25). In the experimental setup for the MVSA-multiple and Multi-ZOL datasets, the corresponding model parameters are set as (0.7, 0.0003, 128, 0.75) and (0.3, 0.0003, 2048, 0.75), respectively.

The dropout value is set as {0.1, 0.3, 0.5, 0.7, 0.9}. Figure 2a shows that, when the dropout is equal to 0.7, we obtain the highest accuracy of 73.17% on the MVSA-single dataset. On the MVSA-multiple dataset, the highest accuracy is 70.45% when the dropout is 0.7. On the Multi-ZOL dataset, the highest accuracy is 74.28%, when the dropout is 0.3.

In Fig. 2b, the learning rate is selected from {0.0001, 0.0002, 0.0003, 0.0005, 0.001}. We observe that all of the three accuracy curves first increase and then decrease. The highest accuracy of 72.73% is obtained when the learning rate equals 0.0003 and 0.0005 on the MVSA-single dataset. The CMCN model exhibits the highest accuracy of 70.45% when the learning rate is 0.0003 on the MVSA-multiple dataset. On the Multi-ZOL dataset, the difference between the highest

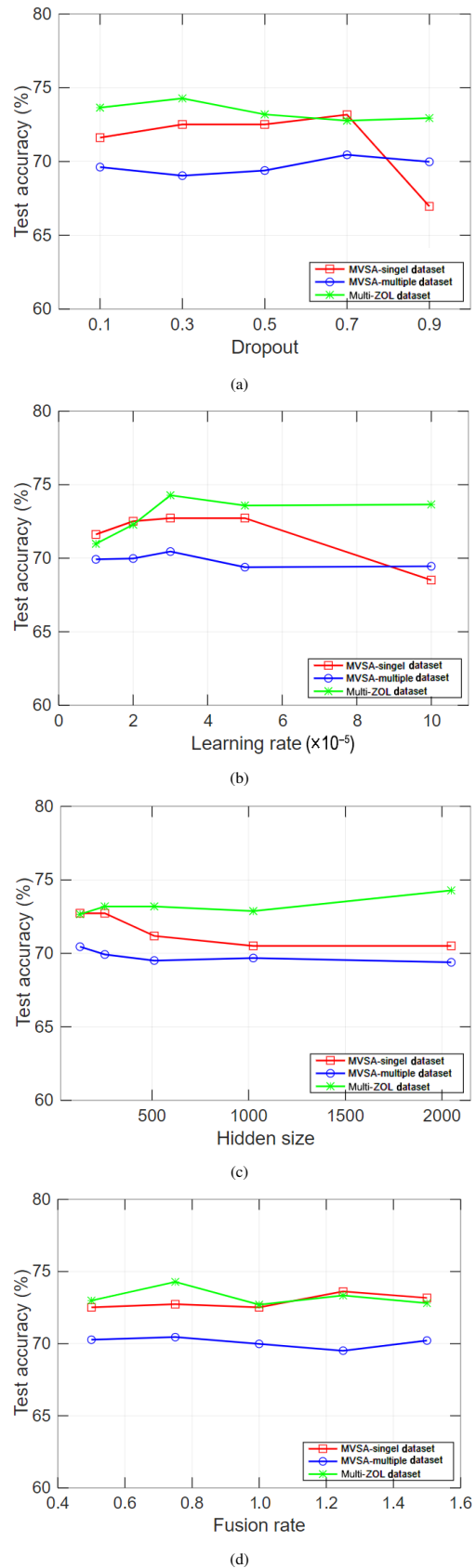


Fig. 2 Parameter experiments on the MVSA-single, MVSA-multiple, and Multi-ZOL datasets.

accuracy of 74.28% and the lowest accuracy of 70.98% on different learning rates is 3.30%.

In Fig. 2c, the size of the hidden layer of the MLP is set as {128, 256, 512, 1024, 2048} to analyze the sensitivity of that hyperparameter. We obtain the best result of 72.73% when the hidden layer size equals 256 on the MVSA-single dataset. The best experimental result of 70.45% is achieved on the MVSA-multiple dataset when the hidden layer size is 128. However, as the hidden layer size increases, the accuracy decreases to 69.39%. When the hidden layer size is equal to 2048, the Multi-ZOL dataset obtains the highest accuracy of 74.28%.

In Fig. 2d, our model learns the fusion ratio {0.5, 0.75, 1.0, 1.25, 1.5} of texts and images during image attention feature extraction. We found that the MVSA-multiple and Multi-ZOL datasets can achieve the best results when the fusion ratio is equal to 0.75. The result of the MVSA-single dataset is the best at 73.61% when the fusion ratio is set as 1.25.

We can conclude from Fig. 2 that the CMCN model is not sensitive to the learning rate, hidden layer, and fusion rate but slightly sensitive to dropout.

#### 4.4.2 Ablation experiments on components

To investigate the contribution of each component to our

CMCN model, we conducted an ablation analysis of each component. The experimental results are summarized in Table 6.

- “Only text” indicates that only text data are used for sentiment classification, whereas “only image” indicates that CNN is used for sentiment classification of image data.
- “Without text-att” indicates that the extracted text attention feature  $F_{tatt}$  is not used in the CMHF module.
- “Without img-att” indicates that the extracted image attention feature  $F_{iatt}$  is not utilized in the CMHF module.
- “Without cross-modal fusion layer” and “Without global fusion layer” indicate that we use the CMHF module, which removes the cross-modal fusion or global fusion layer to fuse features.
- “Without CMHF” discards the first three layers in the CMHF module, directly classifies the sentiment polarity by linear regression for each kind of features (i.e., text features, image features, text attention features, and image attention features), concatenates the classification results, and identifies sentiment polarity on the concatenated classification results.
- “Without fusion rate” indicates that the dot product method is used directly to integrate the images and texts instead of using the hyperparameter  $\theta$  to adjust the ratio

**Table 6** Ablation experiments.

Method	MVSA-single			MVSA-multiple			Multi-ZOL		
	Accuracy	Weighted-F1	Macro-F1	Accuracy	Weighted-F1	Macro-F1	Accuracy	Weighted-F1	Macro-F1
Only text	70.07	71.26	56.82	68.16	68.87	40.06	63.60	64.31	56.55
Only image	67.41	72.40	49.80	67.57	70.69	32.26	54.99	57.77	43.87
Without text-att	70.51	70.87	58.94	68.27	71.77	49.72	74.03	74.05	70.76
Without img-att	72.06	73.74	57.81	68.21	71.03	51.15	73.37	73.45	71.01
Without cross-modal fusion layer	68.74	70.47	53.76	69.33	73.13	50.66	73.16	73.63	69.84
Without global fusion layer	71.18	71.68	58.36	66.75	68.68	49.84	73.54	73.66	70.28
Without CMHF	71.40	74.06	54.04	68.27	69.14	<b>52.64</b>	71.64	71.69	67.34
Without fusion rate	72.73	73.84	59.43	68.57	71.24	49.94	73.09	73.27	69.95
Without corr	72.95	74.45	59.51	68.33	72.88	48.30	72.73	72.80	69.42
Corr by using features directly	72.06	73.16	59.67	68.45	73.14	47.43	73.16	73.36	69.30
Without CMFH + without fusion rate + without corr	72.51	73.43	<b>60.54</b>	67.57	73.71	43.05	69.89	70.16	64.18
Output of feature sampling layer	71.40	73.83	54.87	66.86	68.11	50.98	72.49	72.72	69.18
Output of cross-modal fusion layer	68.96	69.33	59.44	67.63	69.05	52.26	72.91	72.90	69.07
Output of global fusion layer	66.30	66.63	51.70	66.92	68.76	50.00	73.09	73.25	70.60
Final output	43.46	41.76	35.82	54.52	53.36	39.03	73.33	73.65	70.79
CMCN	<b>73.61</b>	<b>75.03</b>	60.51	<b>70.45</b>	<b>74.77</b>	50.58	<b>74.28</b>	<b>74.53</b>	<b>71.51</b>

of image and text fusion.

- “Without corr” indicates that the image-text correlation  $c$  is not used for classification, and “Corr by using features directly” indicates that the model uses the text features extracted by BERT and the image features extracted by VGGNet to calculate the correlation directly.

- “Without CMHF + without fusion rate + without corr” indicates that the first three layers in the CMHF module and image-text correlation  $c$  are not used, and the dot product is used to fuse the image and text directly.

- “Output of feature sampling layer”, “Output of cross-modal fusion layer”, “Output of global fusion layer”, and “Final output” indicate that we take  $y_{pre1}$ ,  $y_{pre2}$ ,  $y_{pre3}$ , and  $y_{pre_{all}}$  in the classification layer as the sentiment polarity for optimization, and analysis. Here,  $y_{pre1}$ ,  $y_{pre2}$ , and  $y_{pre3}$  are the classification results of the feature sampling, cross-modal fusion, and global fusion layers in the CMHF module.  $y_{pre_{all}}$  is the final classification result of the CMCN model.

These results of the ablation experiments show that each component of our CMCN model contributes to its performance.

Table 6 shows that the performance of “Without text-att” or “Without img-att” is lower than that of the CMCN model, which indicates the necessity of using the attention mechanism. Moreover, “Without cross-modal fusion layer”, “Without global fusion layer”, and “Without CMHF” perform worse than the CMCN model, which illustrates the importance of hierarchical fusion for features within a single modality and between two modalities. “Without corr” or “Without fusion rate” also performs worse than the CMCN model. This finding shows that the introduction of image-text correlation can improve classification performance. Furthermore, “Without CMHF + without fusion rate + without corr” has a worse classification result than “Without CMHF”, “Without fusion rate”, and “Without corr”, which demonstrates that these modules can assist each other in optimizing the model.

The CMCN model employs joint optimization, and its performance is better than those of “Output of feature sampling layer”, “Output of cross-modal fusion layer”, “Output of global fusion layer”, and “Final output”. We observe the superiority of the joint optimization method from Table 6. This finding indicates that the joint optimization approach can effectively improve the performance of the CMCN model.

In our CMHF model, the pretraining model BERT and the image feature extraction method VGGNet are only used for feature extraction; they are not involved in the entire model training process. Accordingly, these two parts are not included when calculating the time complexity of the CMHF model. To analyse model complexity, the commonly used image sentiment classification models VGGNet19<sup>[10]</sup> and Resnet101<sup>[44]</sup> are compared. The parameter amount of VGGNet19 is approximately  $1.44 \times 10^8$ , and the parameter amount of Resnet101 is about  $4 \times 10^7$ <sup>\*\*</sup>. By contrast, the parameter amount of our CMHF model is only approximately  $3.3 \times 10^6$ . Thus, our CMHF model is lightweight. Moreover, if the running time of the CMHF model is calculated based on the average of 500 epochs running time on the CPU “i7-9750H”, then the total running time of training and testing for each epoch is about 18 s.

In summary, the carefully designed CMCN technically helps enhance the classification performance, which is hindered by the problem that the correlation between modalities is erroneous or the fusion between modalities is insufficient when learning multimodal data. Extensive experimental results on the three public datasets also show that the proposed approach significantly outperforms the state-of-the-art methods. However, the main disadvantage of the proposed CMCN model is that it needs an image caption generation method to support it. However, currently, image captioning is not a well-solved task in the field of computer vision. In some cases, the descriptive text of the image extracted by the image caption generation method is somewhat not accurate enough. That is, its accuracy may affect the performance of the proposed CMCN model.

## 4.5 Case study




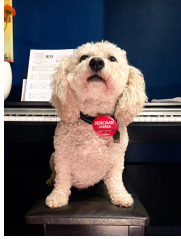
We selected two cases in each of the three datasets for a case study to demonstrate the effectiveness of our CMCN model.

### 4.5.1 Cases about tweets in the MVSA-single and MVSA-multiple datasets

Table 7 shows four cases of multimodal data in the MVSA-single and MVSA-multiple datasets. The feature vectors can be extracted from the texts, images, and descriptive texts, and the image-text correlations can be measured using the text features and descriptive text features.

<sup>\*\*</sup><https://paperswithcode.com/sota/image-classification-on-imagenet>.

**Table 7 Four cases of multimodal data from the MVSA-single and MVSA-multiple datasets.**

	Case (a) from MVSA-single	Case (b) from MVSA-single	Case (c) from MVSA-multiple	Case (d) from MVSA-multiple
Text	RT@CBC: #LoveBirds. Learn about romance in the wild with @cbcdocs #lifestory: #ValentineDay	Quantum of the seas under dramatic clouds. Notice the empty pool– too cold and windy today. Tomorrow Florida.	The Madame Lois - come in and give a try before it sells out! #jamesnorth #hamont #diner #brunch	Max stands tall for @jeromeyc and #RealChange in #Canada #elxn2015 #elxn42 @JustinTrudeau #lpc2015
Image				
Describe text	A couple of birds sitting on top of a rock	A view of a boat in the water	A close up of some food on a table	A small white dog wearing a red bow tie
<b>Without CMFH</b>	Positive	Negative	Neutral	Neutral
<b>Without corr</b>	Positive	Negative	Positive	Positive
<b>CMCN</b>	Positive	Negative	Neutral	Positive
Label	Positive	Negative	Neutral	Positive

Cases (a) and (b) are from the MVSA-single dataset. For Case (a), both the text and image are analyzed to obtain positive sentiments. Thus, the three models “CMCN”, “Without CMHF” and “Without corr” distinguish the sentiment polarity of this tweet as positive. Similar to Case (a), the text and image of Case (b) tend to have negative sentiments. Thus, the three models distinguish the sentiment polarity of this tweet as negative.

Cases (c) and (d) are selected from the MVSA-multiple dataset. In Case (c), the text information does not clearly show the polarity of sentiment, but the image may be detected as having a positive sentiment. Our CMCN model classifies the sentiment of this case as neutral, which is the same as its label. However, “Without corr” does not introduce the relevance of the image and text. Thus, the sentiment polarity of this tweet is judged positive. In Case (d), the classification results of CMCN and “Without corr” are correct, whereas the classification result of “Without CMHF” is incorrect. This incorrect classification result of “Without CMHF” can be attributed to the fact that the neutral features extracted from the text are not fully integrated with the positive features extracted from the image.






#### 4.5.2 Cases about reviews in the Multi-ZOL dataset

Cases (e) and (f) in Table 8 are the reviews of multimodal data in the Multi-ZOL dataset, in which the title and

comment are combined as the original text to extract the original text features. The features extracted from the images are concatenated as the image features. The reviews in the Multi-ZOL dataset contain six aspects, which are “Cost performance”, “Performance configuration”, “Battery life”, “Appearance and feel”, “Camera performance”, and “Screen performance”. In both Cases (e) and (f), only the first five aspects and their corresponding (1–10 points) scores are included in the reviews.

The sentiments in the images of these two cases are not clearly expressed. For Case (e), “Battery life” and “Performance configuration” are highly rated, “Camera performance” is moderately rated, and “Appearance and feel” is poorly rated. Therefore, the result of the CMCN model is accurate in those aspects. For Case (f), “Without CMHF” gives high scores for “Cost performance” and “Battery life”, because it simply concatenates the classification results of the texts and images, with the texts embodying a strong sentimental tendency. “Without corr” does not introduce the association between images and texts, resulting in errors in the aspects “Battery life” and “Appearance and feel”, which are not mentioned in the text. The classification result of our CMCN model is better than those of “Without CMHF” and “Without corr”. Furthermore, our CMCN model improves the two problems of the fusion method, i.e., insufficient and erroneous image-text correlation.

**Table 8 Cases of multimodal review from ZOL.com.**

Title		Case (e) from Multi-ZOL				
(Translated from Chinese)		The biggest disadvantage of this machine is ugliness				
Image						
Comment	(Translated from Chinese) There is nothing bad about the configuration, 960 CPU is very stable now, and the battery capacity is very big too, I can use it for two days without charging, the camera of honor has always been okay. Ugly! Ugly as hell, I don't know who designed those two cameras in the back, the back of the phone looks exactly like R9, come on honor, you can do better. Everything is fine except that it looks a bit ugly.					
Aspect	Performance configuration	Battery life	Appearance and feel	Camera performance	Screen performance	
<b>Without CMFH</b>	8	8	6	6	Null	
<b>Without Corr</b>	8	8	6	6	Null	
<b>CMCN</b>	8	8	6	6	Null	
Label	8	8	8	6	Null	
Title		Case (f) from Multi-ZOL				
(Translated from Chinese)		Samsung S8 – the almost perfect phone				
Image	 (i)	 (ii)	 (ii)	 (iv)		
Comment	(Translated from Chinese) Beautiful look. Running very fast. Sufficient memory. There are still black edges on the side. The system software takes up a lot. I have used it for a while before I comment. Let's talk about the moment I opened the box! Since my original phone is 5.5 inches, I think this phone is too small!!! But it feels good to the hand. I turned on the phone and found that, I don't know who always take images to make it look like there are no black edges no the screen? I am very serious about choosing a mobile phone. For people who always have trouble in choosing, once they determined, they will not choose other ones. Others won't look as good as the one selected! Huawei, Apple, Xiaomi, Honor, Little Green, Little Blue, Smartisan, Meitu, Nubia . . . Basically, whichever phones were released in the last few months, I have seen them thoroughly. And I finally decided to buy this phone that came out long ago! I wanted to buy it when it first came out, but I have been waiting, wanting to see what will happen during the launch events peak season in the second half of the year. But now I discovered that first sight is really important. This phone, except for the black edges on the side, which actually can't be avoided, and only 4G of running memory, everything else is nearly perfect! Fingerprint recognition, facial recognition, iris recognition, except that sometimes I do not know if it is too dark in the bag, or my face does not look like mine, the face unlock would be kind of awkward and I just can't unlock my phone. Luckily you can still unlock the phone by inputting the password, then I quickly recorded my fingerprint to unlock it, really embarrassing! Speaking of this 4G running memory, like many friends said, there are basically just hundreds of megabytes left when you turn on the phone. Fortunately, the system is running smoothly, and there is no forced termination and black screen problem and no redness problem was found.					
Aspect	Cost performance	Performance configuration	Battery life	Appearance and feel	Camera performance	Screen performance
<b>Without CMFH</b>	8	8	8	10	8	Null
<b>Without corr</b>	6	8	10	6	8	Null
<b>CMCN</b>	6	8	6	10	6	Null
Label	6	8	6	10	8	Null

## 5 Conclusion and Future Work

People's comments or reviews on the Internet may contain more than one modal data, and some existing works only consider learning from single-modality data. Moreover, when learning multimodal data, the correlation between modalities could be erroneous, or the fusion between modalities could be insufficient. In this paper, we propose a CMCN with hierarchical fusion for multimodal data of images and texts. Our CMCN model utilizes the image-text correlation generator to reduce the errors caused by integrating images and texts with erroneous correlations. Moreover, the proposed CMHF can integrate the features within a single modality and between two modalities hierarchically. Extensive experimental results show the superiority of this model. In the future, we aim to use graph neural networks combined with hierarchical fusion to address the problem of multimodal sentiment reasoning.

### Acknowledgment

The work was supported by the National Key Research and Development Program of China (No. 2020AAA0104903).

### References

- [1] N. Xu, W. J. Mao, and G. D. Chen, Multi-interactive memory network for aspect based multimodal sentiment analysis, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 371–378, 2019.
- [2] Z. Yu, J. Yu, J. P. Fan, and D. C. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1839–1848.
- [3] Z. Yu, J. Yu, C. C. Xiang, J. P. Fan, and D. C. 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, 2018.
- [4] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, ICON: Interactive conversational memory network for multimodal emotion detection, in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 2594–2604.
- [5] A. Hu and S. Flaxman, Multimodal sentiment analysis to explore the structure of emotions, in *Proc. 24<sup>th</sup> ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 350–358.
- [6] P. Anderson, X. D. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077–6086.
- [7] L. Zhang, S. Wang, and B. Liu, Deep learning for sentiment analysis: A survey, *WIREs Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, 2018.
- [8] S. C. Zhao, S. F. Wang, M. Soleymani, D. Joshi, and Q. Ji, Affective computing for large-scale heterogeneous multimedia data: A survey, *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 3s, p. 93, 2020.
- [9] T. Niu, S. A. Zhu, L. Pang, and A. El Saddik, Sentiment analysis on multi-view social data, in *Proc. 22<sup>nd</sup> Int. Conf. MultiMedia Modeling*, Miami, FL, USA, 2016, pp. 15–27.
- [10] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *Proc. 3<sup>rd</sup> Int. Conf. Learning Representations*, arXiv preprint arXiv:1409.1556v6.
- [11] G. R. Wang, K. Z. Wang, and L. Lin, Adaptively connected neural networks, in *Proc. of the 2019 IEEE Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 1781–1790.
- [12] R. Cadne, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, RUBi: Reducing unimodal biases for visual question answering, in *Proc. 33<sup>rd</sup> Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 841–852.
- [13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
- [14] N. Xu and W. J. Mao, MultiSentiNet: A deep semantic network for multimodal sentiment analysis, in *Proc. 2017 ACM Conf. Information and Knowledge Management*, Singapore, 2017, pp. 2399–2402.
- [15] N. Xu, W. J. Mao, and G. D. Chen, A co-memory network for multimodal sentiment analysis, in *Proc. 41<sup>st</sup> Int. ACM SIGIR Conf. Research & Development in Information Retrieval*, Ann Arbor, MI, USA, 2018, pp. 929–932.
- [16] J. C. Xu, D. L. Chen, X. P. Qiu, and X. J. Huang, Cached long short-term memory neural networks for document-level sentiment classification, in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 1660–1669.
- [17] A. Mishra, K. Dey, and P. Bhattacharyya, Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network, in *Proc. 55<sup>th</sup> Annu. Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 377–387.
- [18] D. H. Ma, S. J. Li, X. D. Zhang, and H. F. Wang, Interactive attention networks for aspect-level sentiment classification, in *Proc. 26<sup>th</sup> Int. Joint Conf. Artificial Intelligence*, Melbourne, Australia, 2017, pp. 4068–4074.
- [19] A. Gaspar and L. A. Alexandre, A multimodal approach to image sentiment analysis, in *Proc. 20<sup>th</sup> Int. Conf. Intelligent Data Engineering and Automated Learning*, Manchester, UK, 2019, pp. 302–309.
- [20] Q. T. Truong and H. W. Lauw, VistaNet: Visual aspect attention network for multimodal sentiment analysis, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 305–312, 2019.
- [21] B. Liu, S. J. Tang, X. J. Sun, Q. Y. Chen, J. X. Cao, J. Z.

- Luo, and S. S. Zhao, Context-aware social media user sentiment analysis, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 528–541, 2020.
- [22] E. J. Barezi and P. Fung, Modality-based factorization for multimodal fusion, in *Proc. 4<sup>th</sup> Workshop on Representation Learning for NLP*, Florence, Italy, 2019, pp. 260–269.
- [23] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, 2018.
- [24] M. H. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L. P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in *Proc. 19<sup>th</sup> ACM Int. Conf. Multimodal Interaction*, Glasgow, UK, 2017, pp. 163–171.
- [25] 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.
- [26] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. B. V. Subramanyam, Benchmarking multimodal sentiment analysis, in *Proc. 18<sup>th</sup> Int. Conf. Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary, 2017, pp. 166–179.
- [27] D. Zhang, S. S. Li, Q. M. Zhu, and G. D. Zhou, Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning, *IEEE Access*, vol. 8, pp. 22945–22954, 2020.
- [28] Z. L. Wang, Z. H. Wan, and X. J. Wan, TransModality: An End2End fusion method with transformer for multimodal sentiment analysis, in *Proc. Web Conf.*, Taipei, China, 2020, pp. 2514–2520.
- [29] C. Yang, X. C. Wang, and B. Jiang, Sentiment enhanced multi-modal Hashtag recommendation for micro-videos, *IEEE Access*, vol. 8, pp. 78252–78264, 2020.
- [30] F. R. Huang, K. M. Wei, J. Weng, and Z. J. Li, Attention-based modality-gated networks for image-text sentiment analysis, *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 16, no. 3, p. 79, 2020.
- [31] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proc. 32<sup>nd</sup> Int. Conf. Machine Learning*, Lille, France, 2015, pp. 2048–2057.
- [32] D. Borth, R. R. Ji, T. Chen, T. Breuel, and S. F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in *Proc. 21<sup>st</sup> ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 223–232.
- [33] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo, A multimodal feature learning approach for sentiment analysis of social network multimedia, *Multimed. Tools Appl.*, vol. 75, no. 5, pp. 2507–2525, 2016.
- [34] G. Y. Cai and B. B. Xia, Convolutional neural networks for multimedia sentiment analysis, in *Proc. 4<sup>th</sup> CCF Conf. Natural Language Processing and Chinese Computing*, Nanchang, China, 2015, pp. 159–167.
- [35] Y. H. Yu, H. F. Lin, J. N. Meng, and Z. H. Zhao, Visual and textual sentiment analysis of a microblog using deep convolutional neural networks, *Algorithms*, vol. 9, no. 2, p. 41, 2016.
- [36] N. Xu, Analyzing multimodal public sentiment based on hierarchical semantic attentional network, in *Proc. 2017 IEEE Int. Conf. Intelligence and Security Informatics*, Beijing, China, 2017, pp. 152–154.
- [37] K. Zhang, Y. S. Geng, J. Zhao, J. X. Liu, and W. X. Li, Sentiment analysis of social media via multimodal feature fusion, *Symmetry*, vol. 12, no. 12, p. 2010, 2020.
- [38] X. C. Yang, S. Feng, D. L. Wang, and Y. F. Zhang, Image-text multimodal emotion classification via multi-view attentional network, *IEEE Trans. Multimed.*, doi: 10.1109/TMM.2020.3035277.
- [39] N. Vo, J. Lu, S. Chen, K. Murphy, and J. Hays, Composing text and image for image retrieval—an empirical odyssey, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6432–6441.
- [40] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, Gated multimodal units for information fusion, in *Proc. 5<sup>th</sup> Int. Conf. Learning Representations*, <https://arxiv.org/abs/1702.01992v1>.
- [41] Y. Q. Wang, M. L. Huang, X. Y. Zhu, and L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 606–615.
- [42] D. Y. Tang, B. Qin, and T. Liu, Aspect level sentiment classification with deep memory network, in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 214–224.
- [43] P. Chen, Z. Q. Sun, L. D. Bing, and W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 452–461.
- [44] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.



**Cheng Peng** received the BEng degree from Beijing Institute of Technology in 2019. He is currently a master student at Beijing Institute of Technology. His research interests include multi-modal sentiment analysis, deep learning, and knowledge graph reasoning.



**Xiaojun Xue** received the BEng degree from China University of Geosciences, China in 2019. He is currently a PhD candidate at Beijing Institute of Technology. His research interests include deep learning and natural language processing.



**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 University of Vermont, USA from January 2010 to February 2011. She is currently an associate professor at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include information extraction, social computing, and machine learning.



**Zhendong Niu** received the PhD degree in computer science from Beijing Institute of Technology, Beijing, China in 1995. From 1996 to 1998, he was a postdoctoral researcher at the University of Pittsburgh, Pittsburgh, PA, USA, where he has been a joint professor at the School of Computing and Information since 2006. He was a researcher and adjunct faculty member at Carnegie Mellon University, Pittsburgh, USA from 1999 to 2004. He is currently a professor at School of Computer Science and Technology, Beijing Institute of Technology. His current research interests include informational retrieval, software architecture, digital libraries, and web-based learning techniques.



**Jiameng Gao** received the BEng degree from Beijing Forestry University, China in 2019. She is currently a master student at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include recommendation system, multi-modal fusion, and sentiment analysis.



**Hongjian Liang** received the BEng degree from Beijing Institute of Technology, China in 2019. He is currently a master student at Waseda University, Japan. His research interests include multimodal sentiment analysis and natural language processing.