Neurocomputing 578 (2024) 127440

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Multi-modal sarcasm detection based on Multi-Channel Enhanced Fusion model

Hong Fang [a,*], Dahao Liang [b], Weiyu Xiang [b]

[a] School of Mathematics, Physics and Statistics, Shanghai Polytechnic University, Shanghai, 201209, Shanghai, China
[b] School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai, 201209, Shanghai, China

ARTICLE INFO

Communicated by D. Cavaliere

*MSC:*
68T01
68T45
68T50

*Keywords:*
Multi-modal sarcasm detection
Attention mechanism
Feature fusion

ABSTRACT

The voluminous quantity of data accessible on social media platforms offers insight into the sentiment disposition of individual users, where multi-modal sarcasm detection is often confounding. Existing sarcasm detection methods use different information fusion methods to combine information from different modalities but ignore hidden information within modalities and inconsistent information between modalities. Discovering the implicit information within the modalities and strengthening the information interaction between modalities is still an important challenge. In this paper, we propose a Multi-Channel Enhanced Fusion (MCEF) model for cross-modal sarcasm detection to maximize the information extraction between different modalities. Specifically, text extracted from images acts as a new modality in the front-end fusion models to augment the utilization of image semantic information. Then, we propose a novel bipolar semantic attention mechanism to uncover the inconsistencies among different modal features. Furthermore, a decision-level fusion strategy from a new perspective is devised based on four models to achieve multi-channel fusion, each with a distinct focus, to leverage their advantages and mitigate the limitations. Extensive experiments demonstrate that our model surpasses current state-of-the-art models in multi-modal sarcasm detection.

## 1. Introduction

Sarcasm is a unique form of language expression in human social activities that can make people express another emotional message instead of their true intentions. Major social platforms, such as Twitter and YouTube, are often filled with satirical remarks in videos, images, and text. Due to the prevalence of sarcasm, extracting semantic information from chaotic multi-modal sarcasm information has become a challenging task. As shown in Fig. 1(a), "looks appetising" in the text shows positive sentiment, while the picture shows an unappetizing pizza. Multi-modal sarcasm detection judges whether there is a sarcasm relationship between them by understanding multi-modal information.

In the field of multi-modal sarcasm detection, the majority of previous research has addressed the problem of diversification of information sources by integrating image and text information. As illustrated in Fig. 1(a), previous researches [1,2] have demonstrated improved performance in addressing the problem of inconsistent emotional expressions between images and text by jointly modeling and integrating information from different modalities. However, these models have shortages in the utilization of image modality information. As shown in Fig. 1(b), it is often challenging to identify sarcasm using only the image and its accompanying text. In such instances, subtitles in the image can have an auxiliary effect on understanding graphic sarcasm. Similarly, as depicted in Fig. 1(c), traditional methods for image feature extraction tend to be inadequate in capturing such information within images. This limitation hinders the precise correlation between text and image, consequently masking sarcasm. Therefore, improving the utilization of image semantic information is a critical issue in solving the sarcasm task. To address this issue, our approach incorporates the use of text from images, alongside region image features, combined with front-end fusion. This strategy is effective in revealing the semantic interpretation of images.

[3] extracts the image attribute features from different image regions using a button-up and top-down attention mechanism [4]. This approach provides additional semantic information for image-based information fusion. By incorporating Graph Convolutional Networks (GCN), the relationship between different modalities is modeled at a fine-grained level, with the attention mechanism employed to highlight cross-modal semantic relationships. Although this method effectively captures semantic information common to both modalities, it struggles with representing semantic information that is diametrically opposed. For example, as shown in Fig. 1(d), the person in the image is drowning, yet gives a thumbs up. In this instance, the "Helps" in the original

* Corresponding author.
*E-mail address:* fanghong@sspu.edu.cn (H. Fang).

https://doi.org/10.1016/j.neucom.2024.127440
Received 17 October 2023; Received in revised form 22 December 2023; Accepted 17 February 2024
Available online 20 February 2024

(a) well that looks appetising ...


(b) when needs to be said


(c) after crappy week i would feel so much better giving one of the things listed to several people.


(d) helps

**Fig. 1.** Sarcastic examples.

text can be more accurately correlated with the drowning area in the image. It is only through the combined interpretation of the opposing semantic signals of the thumbs up and the text in the image that the sarcastic meaning becomes clear. Thus, it becomes essential to simultaneously highlight both the similar and opposing semantic information across modalities to capture the inconsistencies between them. The existence of cross-modal graphs significantly increases the computational complexity of the model. To address this challenge, we design a new bipolar semantic attention mechanism, which strikes a balance between detection efficiency and performance.

In this work, we propose a Multi-Channel Enhanced Fusion (MCEF) model to maximize information extraction and interaction between different modalities. Specifically, to enhance the utilization of image semantic information, we extract text from images as a new modality and integrate this new modality into multiple models through front-end fusion. Furthermore, we propose a new bipolar semantic attention mechanism to reveal inconsistencies between modal features. In addition, given the problem that the models trained by different fusion methods have their characteristics, we designed a decision-level fusion strategy to achieve multi-channel fusion to strengthen the respective advantages of each model and mitigate limitations.

The main contributions of this paper are as follows:

- Text from images is introduced as a new modality to enhance the utilization of image semantic information.
- A new bipolar semantic attention mechanism model is proposed to strengthen the inconsistency between different modal features and improve the detection effect.
- A decision-level fusion strategy is designed to improve the stabilization of multiple models.
- The experimental results on a public multi-modal sarcasm dataset achieve state-of-the-art performance and prove the reliability of our proposed model.

This paper is organized as follows. In Section 1, we provide an introduction to the background of sarcasm detection, including illustrative examples. We outline the main motivation behind our research and present a concise summary of the key points addressed in the paper. Section 2 discusses related work in sarcasm detection, highlighting the differences between previous approaches and our proposed methodology. Section 3 describes our MCEF model for multi-modal sarcasm detection in detail. Section 4 presents the experimental setup, dataset used, and the results of our model, including an ablation study and case study. Finally, Section 5 concludes the paper, summarizing our findings and discussing potential future work in the field of sarcasm detection.

## 2. Related work

### 2.1. Multi-modal sarcasm detection

In recent years, with the rapid development of social media platforms, multi-modal sarcasm detection has gradually become one of the hottest research topics. [1] takes the lead in proposing to use both images and text modalities in the sarcasm detection model. Respectively, this method can easily detect sarcasm on multi-modal datasets, but cannot learn the contextual information between modalities due to the direct splicing of global information. [5] builds a Twitter multi-modal sarcasm dataset and proposes a multi-modal hierarchical fusion model to detect sarcasm. Attributes are extracted from images as the third modality, and representation fusion and modality fusion are used to perform information interaction for classification. However, due to the limitations of image attribute features, some detailed semantic information of some images will still be lost. [6] treats the contradiction between the original text and the hashtags in it as the inconsistency within the modality, and uses the co-attention matrix to model the inconsistency within the modality mold. [3,7] argue that some sarcasm can be expressed through crucial information. GCN is introduced to build interactive in-modal and cross-modal graphs, using text-modality graph to capture contextual information and image-modality graph to capture visual relationships between image blocks. Cross-modal graphs capture different feature information of text-image pairs. This method can effectively establish consistency between modalities but ignores inconsistent information between different modalities, and the graph's structure greatly increases the computational complexity.

In contrast to prior approaches, our research emphasizes the critical role of textual elements within images, such as captions or subtitles, in understanding the semantic context. We implement a bipolar semantic attention mechanism to model the nuanced relationships in sarcastic expressions.

### 2.2. Multi-modal fusion

Multi-modal fusion involves integrating information from multiple modalities, such as text, images, or audio, to enhance system performance [8–12]. It is utilized in various domains, including federated learning, knowledge fusion, image-text matching, and emotion recognition. In federated learning, a unified framework with a co-attention mechanism is developed to fuse complementary information from different modalities, improving the performance of global models [8]. In knowledge fusion, attention-based fusion networks and late fusion methods are employed to align entities from different modalities in knowledge graphs [9,10]. In image-text matching, a multi-view approach with progressive fusion is proposed to leverage inter-modality relationships for better understanding [11]. Finally, in emotion recognition, multi-modal models combine facial expressions, voice tone, speech content, behavior, and physiological features to accurately recognize human emotions [12]. Multi-modal fusion enables the integration of diverse modalities, leading to improved performance and richer representations.

Multi-modal fusion methods can be categorized as: early fusion, interactive fusion, and late fusion. Early fusion is to fuse different modal information before interaction. For example, [5] uses image attribute

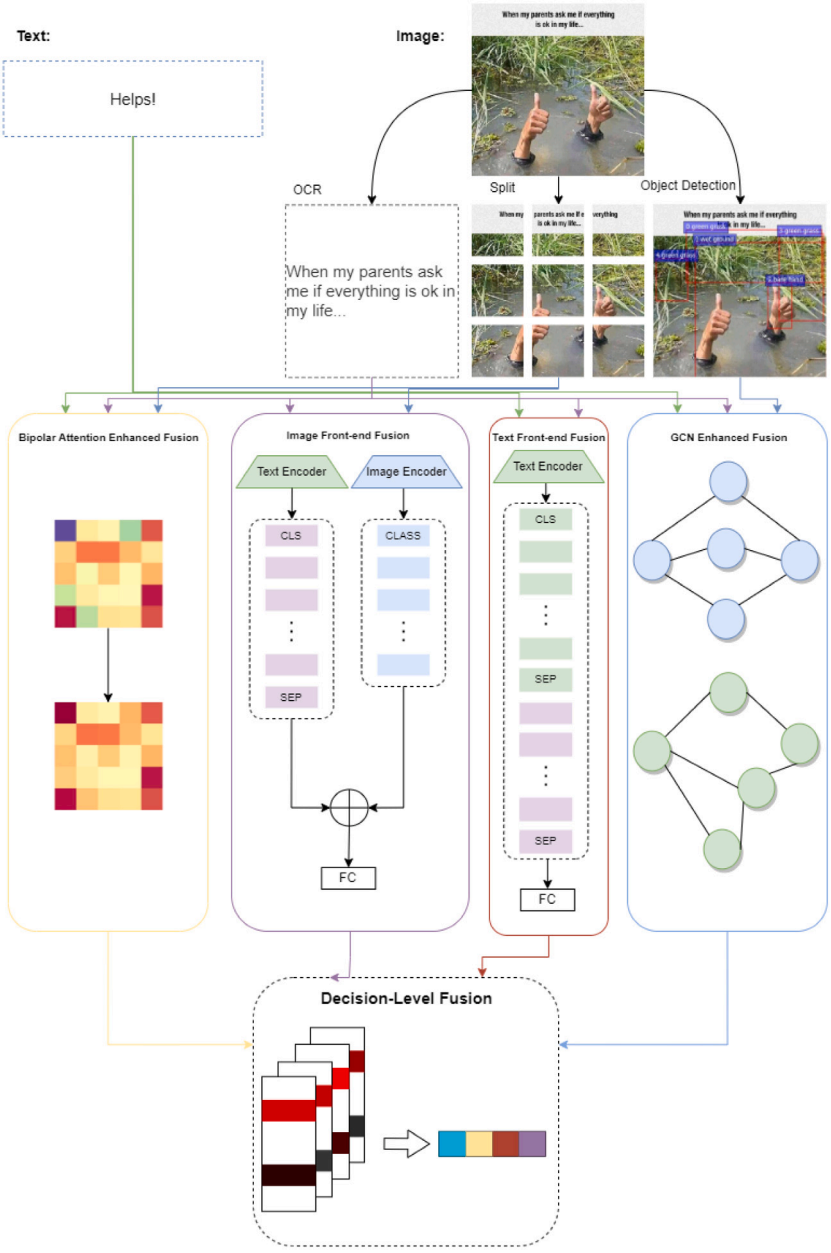



**Fig. 2.** The architecture of the proposed Multi-Channel Enhanced Fusion (MCEF) model framework. Blocks with colored and solid borders represent the models included in MCEF. ⊕ represents matrix concatenation.

features and text for early fusion and then uses Bidirectional Long Short-Term Memory (Bi-LSTM) to extract text features. The interactive fusion focuses on the information interaction between different modalities. For example, [2] uses the feature representation vectors of text and images as the input of the network at the same time and uses Multi-Head Attention with Bidirectional Encoder Representations from Transformers (BERT) to interact with the information between images and text. Late fusion focuses on better representation and utilization of the fusion results. For example, [3,7] utilize graph representations to generate attention scores to weight the fused feature representations. The above studies use different methods to adjust the fused multi-modal feature representation. Still, different types of models also have differences in the key points when detecting sarcasm.

We leverage the textual content extracted from images, employing front-end fusion to enrich the semantic understanding in the early stages of fusion. Additionally, we have formulated a decision-level fusion strategy, a form of late fusion, aimed at augmenting the generalization capabilities and stability when processing diverse types of data. This strategy is meticulously designed to address the unique requirements of different modalities through multiple sub-modules, ensuring a nuanced and comprehensive analysis of each. The decision-level fusion strategy integrates these diverse sub-modules, allowing for a more complete and accurate interpretation of sarcasm.

## 3. Methodology

In this section, we describe our proposed Multi-Channel Enhanced Fusion (MCEF) model for multi-modal sarcasm detection in detail. The overall architecture of the model is shown in Fig. 2. The MCEF model consists of three key components:

- *Feature Extraction and Front-end Fusion*: We extract multi-channel features using Optical Character Recognition (OCR) and Object

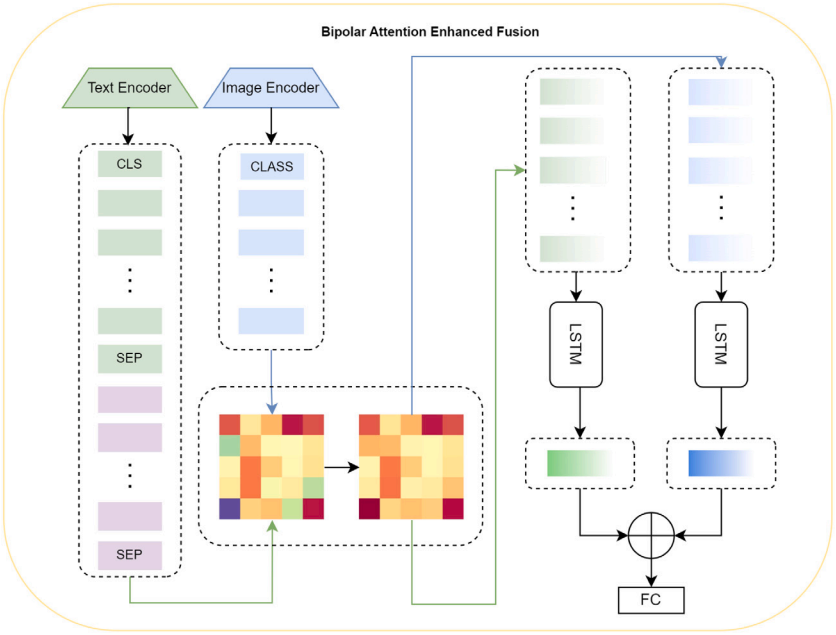



**Fig. 3.** The architecture of the proposed Bipolar Attention Enhanced Fusion (BAEF) model framework.

Detection techniques and treat them as new modalities. The Front-end Fusion strengthens the semantic information within each modality.

- *Enhanced Fusion Models*: We apply a bipolar semantic attention mechanism to discover the similar and opposite semantics between modalities.
- *Decision-level Fusion Strategy*: We balance different models by leveraging their advantages and mitigating limitations, resulting in a more robust sarcasm detection system.

### 3.1. Task definition

Multi-modal sarcasm detection aims to determine whether a multi-modal sample (usually including a text image pair) carries sarcastic intent. Formally, each sample $d$, contains a sequence of words $s = \{w_1, w_2, \ldots, w_n\}$ and an associated image $I$. We aim to learn a multi-modal sarcasm detector that can identify new samples. The detector is represented as a hypothesis $h(x)$ that maps the input sample $d$ to a binary label $y \in \{0, 1\}$, where 1 indicates a sarcastic sample and 0 otherwise.

### 3.2. Feature extraction and front-end fusion

In this section, we provide a detailed overview of our multi-channel information extraction and front-end fusion, aiming to increase the utility of single-modality information and supplement the semantic information of different channels. Our front-end fusion involves using cross-representations to enhance the semantic information of a single modality with semantic information from various channels, and we encourage readers to explore other channels as well. We introduce two sub-modules for our multi-modal sarcasm detection framework: The Image Front-end Fusion (IFF) model and the Text Front-end Fusion (TFF) model. The primary motivation behind these sub-modules is to deepen the semantic understanding within each individual modality. IFF and TFF are designed to primarily focus on single-modality processing, maximizing the potential of each modality in isolation. These models employ an approach of multi-channel information extraction and front-end fusion. This approach significantly amplifies the sarcasm detection capability by intensifying the sarcastic clues when textual content in images is integrated with other modalities.

For image processing, we extract text from images by OCR, treated as a new modality, allowing us to leverage the rich semantic information to enhance our understanding of the visual content. Given an image $I$, we obtain a sequence of words $s^I = \{w_1^I, w_2^I, \ldots, w_{n^I}^I\}$ from the image, where $n^I$ is the length of the text. After adding special tokens `[cls]` and `[sep]`, we map the words into an embedding matrix $X^I = \{x_{cls}^I, x_1^I, \ldots, x_{n^I}^I, x_{sep}^I\}$. To represent the text from images modality, we utilize the pre-trained uncased BERT-base model following [13].

$$T^I = \{t_{cls}^I, t_1^I, t_2^I, \ldots, t_{n^I}^I, t_{sep}^I\} = \text{BERT}(X^I + E_{seg} + E_{pos}) \tag{1}$$

where $T^I \in \mathbb{R}^{(n^I+2)\times d^T}$. $E_{seg} \in \mathbb{R}^{(n^I+2)\times d^T}$ is the segment embedding matrix, $E_{pos} \in \mathbb{R}^{(n^I+2)\times d^T}$ is the embedding matrix.

We resize the image to 224×224 and split it into $p \times p$ patches following the approach of [14]. After adding a special token `[class]`, we map the patches into an embedding matrix $Z^I = \{z_{class}^I, z_1^I, z_2^I, \ldots, z_r^I\}$, where $r = p \times p$. To obtain the image modality representation, we employ the pre-trained Vision Transformer (ViT) model following [14].

$$V = \{v_{class}, v_1, v_2, \ldots, v_r\} = \text{ViT}(Z^I + E_{pos}^I) \tag{2}$$

where $V \in \mathbb{R}^{(r+1)\times d^I}$, and $E_{pos}^I \in \mathbb{R}^{(r+1)\times d^I}$ is the positional embedding matrix.

To integrate the text from images representation and the image representation, we propose the *Front-end Fusion*. Specifically, we concatenate the `[class]` token representation $v_{class}$ and the `[cls]` token representation $t_{cls}^I$ to obtain the front-end fused feature $f_1$. The **Image Front-end Fusion** (IFF) model is then formed by feeding $f_1$ into a fully connected network to make predictions.

$$f_1 = \{v_{class}, t_{cls}^I\} \tag{3}$$

$$\hat{y}_1 = \text{softmax}(W_1 f_1 + b_1) \tag{4}$$

For text processing, given a sequence of text $s = \{w_1, w_2, \ldots, w_n\}$, where $n$ is the length of the text, we adopt a different *Front-end Fusion* for the text modality compared to the image modality. Specifically, we use the BERT model to get the text modality representation $T^{IT}$ as follows:

$$\begin{aligned} T^{IT} &= \{t_{cls}^{IT}, t_1^I, t_2^I, \ldots, t_{n^I}^I, t_{sep}^{IT}, t_1, t_2, \ldots, t_n, t_{sep}^{IT}\} \\ &= \text{BERT}(X^{IT} + E_{seg} + E_{pos}) \end{aligned} \tag{5}$$

where $T^{IT} \in \mathbb{R}^{(n^I+n+3)\times d^T}$. The special token [sep] split the two sentences. Bert handles two sentences with segment embedding mechanism. We take the [cls] token representation as the front-end fused feature to predict, as we call **Text Front-end Fusion** (TFF) model.

$$f_2 = \{t^{TI}_{cls}\} \tag{6}$$

$$\hat{y}_2 = \mathrm{softmax}(W_2 f_2 + b_2) \tag{7}$$

### 3.3. Enhanced fusion models

In this section, we focus on relation extraction of multi-modal semantic nodes to fit the multi-modal sarcasm detection task. We propose a *Bipolar Attention Mechanism*, aiming at retaining both identical and opposite semantic information and model multi-modal semantic relation (see Fig. 3). Our approach has a smaller computational complexity compared to GCN-based approaches, which construct an in-modal and cross-modal graph to model the relationships. Building upon our framework, we introduce two critical sub-modules: the GCN Enhanced Fusion (GEF) model and the Bipolar Attention Enhanced Fusion (BAEF) model. Both sub-modules are engineered with a focus on semantic interaction in a multi-modal context, aimed at uncovering the sarcastic expressions. The GEF model utilizes the strengths of GCN to facilitate a deeper integration of multi-modal information. By leveraging GCN, GEF can effectively map and understand the complex interaction of semantic elements across different modalities. Concurrently, the BAEF model is designed around the innovative Bipolar Attention Mechanism. By focusing on these bipolar semantic aspects, BAEF excels in identifying subtle indicators of sarcasm that might otherwise be overlooked. The goal of both GEF and BAEF is to adapt to a wide range of semantic inputs from multi-channel features.

Firstly, we employ the front-end fused features $V = \{v_{class}, v_1, v_2, \dots, v_r\}$ and $T^{TI} = \{t^{TI}_{cls}, t_1, t_2, \dots, t_n, t^{TI}_{sep}, t^I_1, t^I_2, \dots, t^I_{n^I}, t^{TI}_{sep}\}$, where the text at the first sentence and the text from the image at the second sentence. $T^{TI} \in \mathbb{R}^{(n+n^I+3)\times d^T}$ and $V \in \mathbb{R}^{(r+1)\times d^I}$, and the dimension of the representations is $d^e = d^T = d^I$. We apply a simplified scaled dot-product attention between these feature vectors:

$$S = \frac{T^{TI} V^\top}{\sqrt{d^e}} \tag{8}$$

Next, we apply the *Bipolar Attention Mechanism* to activate negative similarity nodes and obtain attention vectors, as shown below:

$$T^b = T^{TI}\,\mathrm{softmax}(|S|) \tag{9}$$

$$V^b = V\,\mathrm{softmax}(|S^\top|) \tag{10}$$

where $T^b \in \mathbb{R}^{(n^r+1)\times d^e}$ represents the bipolar text modality features, and $V^b \in \mathbb{R}^{(n+n^I+3)\times d^e}$ represents the bipolar image modality features.

We then use Bi-LSTM to aggregate attention vectors, concatenate the output of both, and use a Multi-Layer Perceptron (MLP) layer as the classification layer. This model is called the **Bipolar Attention Enhanced Fusion** (BAEF) model, as shown below:

$$T^f = \{t^f_1, \dots, t^f_{(r+1)}\} = \mathrm{BiLSTM}(T^b) \tag{11}$$

$$V^f = \{v^f_1, \dots, v^f_{(n+n^I+3)}\} = \mathrm{BiLSTM}(V^b) \tag{12}$$

$$f_3 = \{t^f_1, v^f_1\} \tag{13}$$

$$\hat{y}_3 = \mathrm{softmax}(W^2_3 \mathrm{Relu}(W^1_3 f_3 + b^1_3) + b^2_3) \tag{14}$$

where LSTM has a hidden dimension $d^h$, $T^f \in \mathbb{R}^{(r+1)\times d^{2h}}$, $V^f \in \mathbb{R}^{(n+n^I+3)\times d^{2h}}$.

Moreover, aiming to capture the semantic relationships among nodes, we leverage a bottom-up attention model following [4] for object detection. We process the object regions using the ViT encoder, which yields the representation $V^o = \{v^o_1, v^o_2, \dots, v^o_{n^o}\}$, where $n^o$ is the number of regions obtained by the bottom-up attention model, and $v^o_i \in \mathbb{R}^{d^I}$. To learn the incongruity relations of within and across modalities, we construct a graph convolution network (GCN) with the front-end fused features $T^{TI}$ and $V^o$. In brief, an in-modal graph is constructed by the dependency tree of the sentence, and a cross-modal graph is constructed by the sentiment relation similarity between image regions and text tokens, following the approach in [3]. We refer to this model as the **GCN Enhanced Fusion** (GEF) model. The following formula to define the graph convolution computation:

$$X^{l+1} = \sigma\left(\tilde{A} X^l W^l + b^l\right) \tag{15}$$

Here, $X^l$ represents the layer-wise multi-channel feature representations, which in the context of our GEF model, correspond to the front-end fused features, namely $T^{TI}$ and $V^o$. The adjacency matrix $A$ is constructed based on the sentiment relation similarity between image regions and text tokens, as outlined in [3]. The matrix $\tilde{A}$, which is the normalized symmetric adjacency matrix, is computed from $A$. GCN allows for the processing of the relational information in both the text and image modalities, facilitating a more nuanced understanding of the incongruity relations within and across these modalities in the GEF model.

### 3.4. Decision-level fusion strategy

In this section, we propose a *Decision-level Fusion Strategy* to address potential semantic deficiencies in models that use diverse channels and modalities. This strategy also leverages the confidence of different models to exploit each model's strengths. We consider four models: *Image Front-end Fusion*, *Text Front-end Fusion*, *GCN Enhanced Fusion*, and *Bipolar Attention Enhanced Fusion*, each of which employs a different fusion approach to incorporate multi-channel modal information and has a unique advantage for sarcasm detection. Image and text front-end fusion models offer deeper semantics from representations themselves. Bipolar attention enhanced fusion model focuses on the semantics of opposites and approximations, while GCN enhanced fusion model learns the relationship from in-modal and cross-modal graphs. The models above are trained by minimizing the cross-entropy loss given the ground-truth label $y$ and prediction $\hat{y}$. All models are trained separately.

We collect the logits outputs from each trained model and normalize them. We then assign a weight to each model output based on a grid search on the validation set. The normalized logits are fed into a softmax function, and the resulting probabilities are used for prediction.

Let $Z^l = \{z^l_1, z^l_2, \dots, z^l_m\}$ be the logits outputs of $m$ models, where $z^l_i$ is the logits of the $i$th model. The decision-level fusion output $\hat{y}$ is computed as:

$$\hat{y} = \mathrm{softmax}(W_d \mathrm{norm}(Z^l)) \tag{16}$$

where $W_d$ is the weights for decision-level fusion strategy on multiple models, and norm denotes a normalization operation.

## 4. Experiment and results

### 4.1. Experimental dataset

In our study, we utilize the Twitter multi-modal sarcasm detection dataset collected by [5], which is developed specifically for evaluating the multi-modal sarcasm detection task.[1] The dataset comprises English tweets that contain a picture. The final dataset consists of over 24,000 sentences, with roughly equal proportions of positive and negative examples. The dataset is divided into three subsets with an 80:10:10 ratio, namely, training, development, and testing. Table 1 presents the statistics of the dataset.

[1] https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

**Table 1**
Statistics of the dataset.

| | Training | Development | Testing |
|---|---|---|---|
| Sentences | 19 816 | 2410 | 2409 |
| Positive | 8642 | 959 | 959 |
| Negative | 11 174 | 1451 | 1450 |

**Table 2**
Hyper-parameters.

| Hyper-parameters | IFF | TFF | GEF | BAEF | MCEF |
|---|---|---|---|---|---|
| Maximum sequence length | 100 | 100 | 100 | 100 | – |
| Patch size | 16 | 16 | – | 16 | – |
| Number of visual regions | – | – | 10 | – | – |
| $d^T$ | 768 | 768 | 768 | 768 | – |
| $d^I$ | 768 | – | 768 | 768 | – |
| Learning rate | 1e−2 | 1e−3 | 2e−5 | 1e−3 | – |
| Learning rate for fine-tuning | 2e−5 | 2e−5 | 2e−5 | 2e−5 | – |
| Weight decay | 1e−5 | 1e−5 | 1e−5 | 1e−5 | – |
| Batch size | 32 | 32 | 32 | 16 | – |
| $d^h$ | – | – | 512 | 512 | – |
| Number of bipolar attention heads | – | – | – | 2 | – |

### 4.2. Experimental settings

**Preprocessing.** NLTK and spaCy toolkits are utilized to preprocess text and derive dependency trees. Object detection is performed following the approach presented in [4]. OCR is conducted using Google's Tesseract-OCR Engine, and the resulting text is cleaned. For the image front-end fusion model, an image augmentation of random resize and random horizontal flip are applied.

**Pre-trained models.** The pre-trained uncased BERT-base model and ViT-base model are used to obtain text and image modality representations.

**Optimization.** The optimizer for all models is Adam.

**Hyper-parameters.** The GCN enhanced fusion model is utilized, and the settings to construct the adjacency matrix are mostly the same as in [3]. The bipolar attention enhanced fusion model is also utilized, and the batch size is set to 16. Other hyper-parameters are listed in Table 2.

### 4.3. Baselines

To measure the effectiveness of our proposed MCEF model, we use Accuracy, Precision, Recall, and F1-score as performance metrics. In order to compare our model to existing state-of-the-art models, we evaluate the following:

*Image-modality methods*: Models that use only visual information.

- **ResNet**: following the work of [5], use pre-trained ResNet [15] and only update the classification layer parameters.
- **ViT**: using the pre-trained ViT [16] [class] token for sarcasm detection.
- **IFF**: our Image Front-end Fusion model.

*Text-modality methods*: Models that use only textual information.

- **TextCNN**: [17] uses a CNN to classify sarcasm on text.
- **Bi-LSTM**: uses a bi-directional LSTM to classify sarcasm on text.
- **SIARN**: [18] employs inner-attention for textual sarcasm detection.
- **SMSD**: [19] explores a self-matching network to capture textual incongruity information.
- **BERT**: uses pre-trained BERT-base [13] for sarcasm detection on text.
- **TFF**: our Text Front-end Fusion model.

*Multi-modal methods*: Models that take both text and image modality information.

- **HFM**: [5] proposes a hierarchical fusion approach that takes image features, image attribute features, and text features as input.
- **D&R Net**: [20] models both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network.
- **Res-BERT**: [6] concatenates the image features and the text feature for sarcasm prediction based on BERT.
- **Att-BERT**: [6] explores inter-modality attention to model the incongruity between modalities and co-attention mechanisms to model the incongruity within text modality based on BERT.
- **InCrossMGs**: [7] establishes in-modal and cross-modal graphs, models semantic node relationships.
- **CMGCN**: [3] applys object detection to enhance the image region and builds intra-modality and inter-modality graphs.
- **GEF**: our GCN Enhanced Fusion model.
- **BAEF**: our Bipolar Attention Enhanced Fusion model.

**Table 3**
Main experimental results. Results of baselines with the †are retrieved from [3].

| Modality | Models | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| V | ResNet† | 64.76 | 54.41 | 70.80 | 61.53 |
| | ViT† | 67.83 | 57.93 | 70.07 | 63.43 |
| | IFF (ours) | 74.43 | 69.86 | 68.51 | 67.20 |
| T | TextCNN† | 80.03 | 74.29 | 76.39 | 75.32 |
| | Bi-LSTM† | 81.90 | 76.66 | 78.42 | 77.53 |
| | SIARN† | 80.57 | 75.55 | 75.70 | 75.63 |
| | SMSD† | 80.90 | 76.46 | 75.18 | 75.82 |
| | BERT† | 83.85 | 78.72 | 82.27 | 80.22 |
| | TFF (ours) | 85.31 | 80.96 | 82.48 | 81.72 |
| T+V | HFM† | 83.44 | 76.57 | 84.15 | 80.18 |
| | D&R Net† | 84.02 | 77.97 | 83.42 | 80.60 |
| | Res-BERT† | 84.80 | 78.87 | 84.46 | 81.57 |
| | att-BERT† | 86.05 | 80.87 | 85.08 | 82.92 |
| | InCrossMGs† | 86.10 | 81.38 | 84.36 | 82.84 |
| | CMGCN† | 87.55 | 83.63 | 84.69 | 84.16 |
| | GEF (ours) | 84.77 | 81.42 | 79.98 | 80.69 |
| | BAEF (ours) | 85.84 | 81.79 | 82.90 | 82.34 |
| | **MCEF (ours)** | **87.80** | **84.10** | **85.50** | **84.80** |

### 4.4. Main results

In this section, we test our model against several existing baselines. We report the comparison results in Table 3. Our proposed MCEF model outperforms all other models across all metrics, achieving an accuracy of 87.80% and an F1-score of 84.80%.

We find that our BAEF model performed particularly well, achieving an F1-score of 82.34%, which is higher than all baselines excluding GCN-based models, which suffer from much higher computational complexity. We also test our model on single-modal data and find that our Front-end Fusion models, which introduce a new modality of text from images, show significant improvements. These models leverage the semantic information contained in images, even when some images do not explicitly include text, to improve sarcasm detection accuracy. Specifically, our TFF model achieves an accuracy of 85.31% and an F1-score of 81.72%, which is significantly higher than other text-based models. Moreover, taking the outputs of the four models using a Decision-level Fusion Strategy can effectively address the limitations of individual models in terms of insufficient semantic information and the interference of different modal information on sarcasm detection results. This is demonstrated by our MCEF model, which utilizes the Decision-level Fusion Strategy to significantly improve the stability, resulting in better performance than other baselines.

These results demonstrate the effectiveness of our MCEF model, particularly when combined with the muti-channel semantic information contained in images. They also highlight the potential of our Front-end Fusion models to improve sarcasm detection accuracy across all modalities.

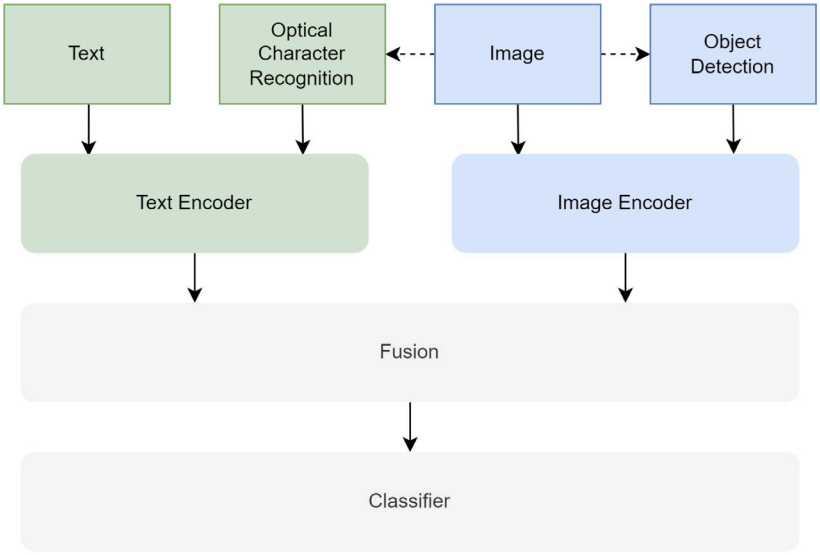



**Fig. 4.** The architecture of the proposed model for sentiment analysis task.

**Table 4**
Results on MVSA-Multiple dataset. Results with †are retrieved from [21].

| Models | Accuracy | F1-score |
|---|---|---|
| MultiSentiNet [22]† | 68.86 | 68.11 |
| HSAN [23]† | 67.96 | 67.76 |
| Co-MN-Hop6[24]† | 68.92 | 68.83 |
| CLMLF [25]† | 71.12 | 68.63 |
| MVCN [21]† | 72.07 | 70.01 |
| Ours | 71.88 | 67.87 |

**Table 5**
Ablation study results. All results are from our experiments.

| Models | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| I Base | 72.31 | 64.69 | 67.05 | 65.85 |
| I Base + IFF | 74.43 | 69.86 | 68.51 | 67.20 |
| T Base | 83.94 | 79.01 | 81.23 | 80.10 |
| T Base + TFF | 85.31 | 80.96 | 82.48 | 81.72 |
| TFF + GEF | 84.77 | 81.42 | 79.98 | 80.69 |
| TFF + BAEF | 85.84 | 81.79 | 82.90 | 82.34 |
| TFF + AEF | 85.14 | 80.26 | 83.11 | 81.66 |
| MCEF (ours) | 87.80 | 84.10 | 85.50 | 84.80 |

Currently, the availability of datasets that encompass both text and images for sarcasm detection is quite limited. Our experiments on the MVSA-Multiple dataset serve as a study. The dataset is designed for multi-modal sentiment analysis [26]. The reason for selecting the MVSA-Multiple dataset is the commonalities shared between sarcasm detection and sentiment analysis tasks. Both tasks necessitate a deep understanding of semantics, as they rely on the ability to interpret and integrate complex semantic information across different modalities. By employing our multi-channel Feature Extraction and Front-end Fusion techniques, we aim to delve deeper into the modal semantic information extraction and understanding. In the experiment, OCR and Object Detection were utilized to extract text from images and focus on the region semantic information in images, shown in Fig. 4. With Front-end Fusion, we could find that the proposed method can also extract more semantic information from both modalities in multi-modal sentiment analysis tasks. The comparison of experimental results can be seen in Table 4. However, there are differences in the semantic information emphasized by different multi-modal tasks, so the experimental results are not the best. We will further investigate this task in the future.

### 4.5. Ablation study

In this section, we aim to investigate the impact of different modalities and fusion strategies on the performance of our multi-modal sarcasm detection model. Table 5 summarizes the accuracy and F1-score of various models.

To start with, we fine-tune the ViT-base model on the image data and achieve accuracy as an Image-base model (I Base). We then perform Image Front-end Fusion (IFF) between the image and text from images modalities, which yield a similar F1-score of 67.20%. For the text modality, we fine-tune the BERT-base model as a Text-base model (T Base). Applying Text Front-end Fusion (TFF) with the text from images modality further improves the accuracy and F1-score to 85.31% and 81.72%, respectively. These results suggest that adding the semantic information of the text from images as a new modality significantly improves the utilization of image semantics and compensates for the deficiencies of other modalities.

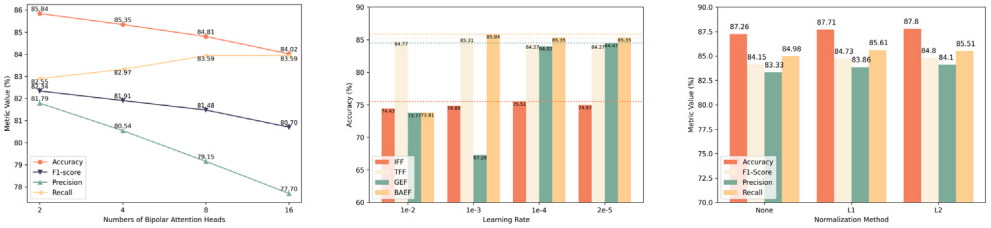
We also investigate the impact of different fusion strategies on the model's performance. We apply the GCN Enhanced Fusion (GEF) and get an F1-score of 80.69%, which does not reach the same results as GMGCN. The Bipolar Attention Enhanced Fusion (BAEF), which uses a smaller computational complexity than the GEF, achieves a better performance with an accuracy of 85.84% and an F1-score of 82.34%. The Attention Enhanced Fusion (AEF), without the bipolar operation, yields a slightly lower result.

Additionally, we conduct experiments to investigate the impact of introducing new modalities on the model's performance. While adding multi-channel features can improve results for some samples, it can also interfere with the detection results of the original samples. To address this conflict, we take the outputs of the four models using a decision-level fusion strategy, i.e., MCEF, which significantly improves the accuracy and F1-score of our model.

Our ablation study highlights the effectiveness of text from images as a new modality, front-end fusion, bipolar semantic attention mechanism, and decision-level fusion strategy in improving the performance of our multi-modal sarcasm detection model. Our findings also emphasize the potential impact of introducing new channels and modalities on the performance of multi-modal models.

### 4.6. Analysis

We conduct an analysis of the effect of varying the number of bipolar attention heads on the performance of our BAEF model. We




(a) BAEF performance of using different numbers of bipolar attention heads

(b) Performance of using different learning rate on 4 models

(c) MCEF performance of using different normalization

**Fig. 5.** Visualization of parameters analysis.

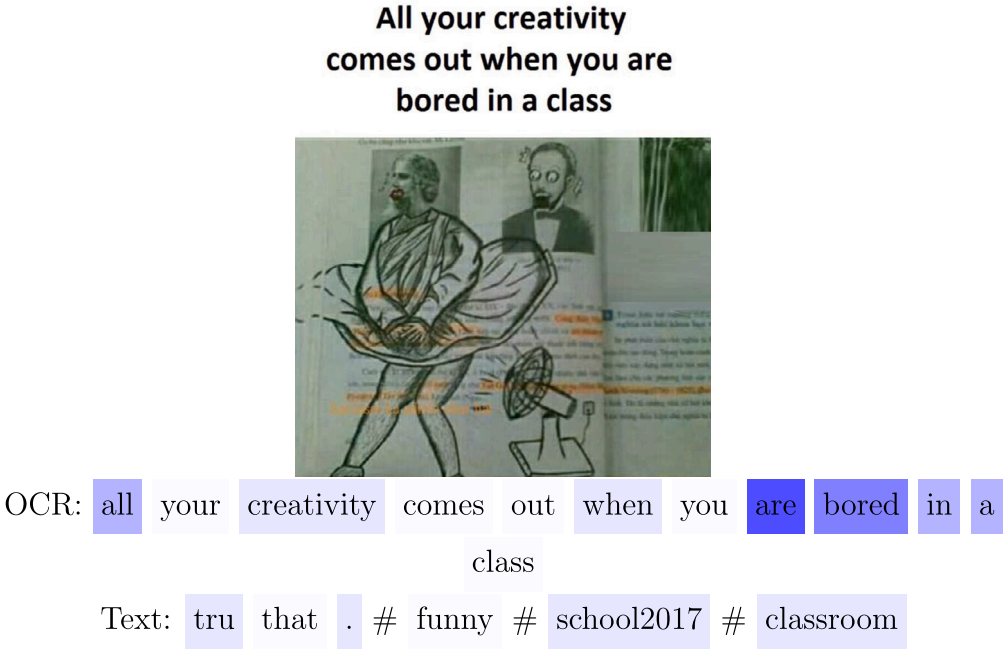



**Fig. 6.** Analysis of a typical example.

vary the number of attention heads from 2 to 16 and present the results in Fig. 5(a). Our findings indicate that increasing the number of attention heads beyond 2 leads to a decline in model performance, possibly due to the loss of relevant features. We also conduct an analysis of the impact of learning rate on 4 models in Fig. 5(b), which helps us to tune the parameters. The impact of different normalization methods of decision-level fusion strategy is shown in Fig. 5(c). We can find that MCEF remains stable in different normalization methods with decision-level fusion strategy.

We visualize an example of front-end fusion using the text from images as a new modality. Some of the first layer attention weights from the text encoder are shown in Fig. 6. If the key cues (marked in blue) are captured in text form images, then the correct label for this example is easily inferred. This demonstrates the effectiveness of the text from images and front-end fusion.

In Table 6, several examples of sarcasm detection are shown. In our case studies, several error types commonly emerge. Firstly, when semantic understanding is primarily derived from text within images, traditional image models often fail to interpret this textual content, as seen in Table 6(a) and (d). This highlights the necessity for models to process both visual elements and text extracted from images. Secondly, errors can occur when image features are not distinctly recognizable, as exemplified in Table 6(b). This highlights the challenge of revealing clearer visual elements. Thirdly, errors often arise from an inability to fuse text and image modalities, as shown in Table 6(c). This lack of comprehensive fusion limits the understanding of the sarcastic meaning, indicating a need for more advanced fusion techniques. Lastly, the presence of misleading elements in either the textual or visual modality can lead to misinterpretations, as demonstrated in Table 6(e). These error types highlight the complex nature of multi-modal sarcasm detection and the necessity for robust models capable. However, we still encounter challenges with the example in Table 6(e). We conjecture that there are three reasons for this failure:

- The associated semantics in the text are ambiguous, focusing solely on “the show”.
- The details of the dog’s face are difficult to recognize in the image, possibly due to the color of the dog.
- There is an evident semantic conflict between the dog’s face and the abstract facial expressions of humans in this image,

**Table 6**
Examples of multi-modal sarcasm detection with the proposed model.

| | Text | Image | Ground Truth | GEF | MCEF |
|---|---|---|---|---|---|
| (a) | good one panda. (insert) |  | positive | negative | positive |
| (b) | dad's handy work. cannot tell at all. |  | positive | positive | positive |
| (c) | snowy owl came too close to traffic camera and he is being ticketed for not reading the signs in # french |  | positive | positive | positive |
| (d) | lunch time emoji_19 |  | positive | negative | positive |
| (e) | my dogs face looks like she just saw the best broadway show |  | positive | negative | negative |

which can mislead the model in understanding the true semantic information.

## 5. Conclusion and future work

In this work, we propose a multi-modal sarcasm detection model MCEF, which takes advantage of multi-channel features from multiple modalities. We introduce a new modality of text extracted from images and utilize front-end fusion, enhanced fusion, and decision-level fusion strategies to effectively combine information from multiple models. Experimental results on a public dataset demonstrate the effectiveness and superiority of our proposed model, which outperforms recent state-of-the-art approaches. Moreover, our front-end fusion models, which incorporate text from images, show significant improvement over other models in single-modal data experiments. These results suggest that the multi-channel features from multiple modalities, as well as the use of different fusion strategies, can lead to improving performance in sarcasm detection.

In this paper, we adopt a multi-channel enhanced fusion strategy. However, more channels are yet to be discovered. If there is an end-to-end model that can easily include more channels, even with external knowledge, with a more general fusion approach, which makes it no longer necessary to change the models for the newly joint channel information. In addition, the multi-channel approach can be migrated to other multi-modal tasks, which focus on the efficient information fusion of texts and images.

## CRediT authorship contribution statement

**Hong Fang:** Formal analysis, Funding acquisition, Writing – review & editing, Conceptualization, Project administration, Supervision, Methodology, Writing – original draft. **Dahao Liang:** Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Weiyu Xiang:** Data curation, Investigation, Methodology, Resources, Validation, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgment

This work is supported by China University Industry-Research Innovation Fund for Next-Generation Information Technology Innovation Projects (2021ITA03008).

## References

[1] R. Schifanella, P. de Juan, J.R. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, the Netherlands, October 15-19, 2016, ACM, 2016, pp. 1136–1145.
[2] X. Wang, X. Sun, T. Yang, H. Wang, Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data, in: Proceedings of the First International Workshop on Natural Language Processing beyond Text, Association for Computational Linguistics, Online, 2020, pp. 19–29.
[3] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 1767–1777.
[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6077–6086.
[5] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in Twitter with hierarchical fusion model, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 2506–2515.
[6] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, EMNLP 2020, in: Findings of ACL, Association for Computational Linguistics, 2020, pp. 1383–1392.
[7] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, R. Xu, Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, ACM, 2021, pp. 4707–4715.
[8] B. Xiong, X. Yang, F. Qi, C. Xu, A unified framework for multi-modal federated learning, Neurocomputing 480 (2022) 110–118.
[9] B. Cheng, J. Zhu, M. Guo, MultiJAF: Multi-modal joint entity alignment framework for multi-modal knowledge graph, Neurocomputing 500 (2022) 581–591.
[10] H. Guo, J. Tang, W. Zeng, X. Zhao, L. Liu, Multi-modal entity alignment in hyperbolic space, Neurocomputing 461 (2021) 598–607.
[11] J. Wu, L. Wang, C. Chen, J. Lu, C. Wu, Multi-view inter-modality representation with progressive fusion for image-text matching, Neurocomputing 535 (2023) 1–12.
[12] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing 561 (2023) 126866.
[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.
[14] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, M. Tomizuka, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer visionCoRR abs/2006.03677, 2020.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[17] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1746–1751.

[18] Y. Tay, A.T. Luu, S.C. Hui, J. Su, Reasoning with sarcasm by reading in-between, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 1010–1020.

[19] T. Xiong, P. Zhang, H. Zhu, Y. Yang, Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 2115–2124.

[20] N. Xu, Z. Zeng, W. Mao, Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 3777–3786.

[21] Y. Wei, S. Yuan, R. Yang, L. Shen, Z. Li, L. Wang, M. Chen, Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 5240–5252.

[22] N. Xu, W. Mao, MultiSentiNet: A deep semantic network for multimodal sentiment analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017, ACM, 2017, pp. 2399–2402.

[23] N. Xu, Analyzing multimodal public sentiment based on hierarchical semantic attentional network, in: 2017 IEEE International Conference on Intelligence and Security Informatics, ISI 2017, Beijing, China, July 22-24, 2017, IEEE, 2017, pp. 152–154.

[24] N. Xu, W. Mao, G. Chen, A co-memory network for multimodal sentiment analysis, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 929–932.

[25] Z. Li, B. Xu, C. Zhu, T. Zhao, CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection, in: Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 2282–2294.

[26] T. Niu, S. Zhu, L. Pang, A. El-Saddik, Sentiment analysis on multi-view social data, in: MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 9517, Springer, 2016, pp. 15–27.

**Dr. Hong Fang** is a member of the School of Mathematics, Physics and Statistics at Shanghai Polytechnic University. After earning her Ph.D. in Computer Science and Technology from Anhui University in 2008, Dr. Fang has been instrumental in advancing the fields of Natural Language Processing and Data Mining. Over her career, Dr. Fang has published more than 20 academic papers, contributing significantly to the knowledge in her areas of expertise. fanghong@sspu.edu.cn.

**Dahao Liang** received the B.S. degree in Donghua University in 2019. He is currently working toward the M.S. degree at the School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University. Liang's research focuses on multimodal tasks and adversarial examples.

**Weiyu Xiang** received the B.S. degree in Shanghai Polytechnic University in 2021. He is currently working toward the M.S. degree at the School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University. Xiang's research primarily focuses on multimodal tasks and object detection.