Exploring Multimodal Multiscale Features for Sentiment Analysis Using Fuzzy-Deep Neural Network Learning

Xin Wang, Jianhui Lyu, Byung-Gyu Kim, *Senior Member, IEEE*, B. D. Parameshachari, Keqin Li, *Fellow, IEEE*, and Qing Li, *Senior Member, IEEE*

***Abstract*—Sentiment analysis, a challenging task in understanding human emotions expressed through diverse modalities, prompts the development of innovative solutions. Multimodal data often contains important complementary information. Effective fusion and extraction of multimodal data features are key issues in sentiment analysis. In this article, we introduce a novel sentiment analysis model that integrates multimodal multiscale features based on a fuzzy-deep neural network. First, we combine multimodal data, namely text, audio, and images, to extract intrinsic feature representations. Second, our model incorporates the fuzzy-deep neural network learning module, infused with fuzzy logic principles to enhance adaptability to the inherent vagueness in sentiment expressions. Furthermore, we integrate the dual attention mechanism that dynamically focuses on pivotal aspects within multimodal data, refining feature extraction for heightened context-awareness. Rigorous validation across three datasets, including the Multimodal Corpus of Sentiment Intensity dataset, the Multimodal Opinion Sentiment and Emotion Intensity dataset, and the Chinese Single and Multimodal Sentiment dataset, demonstrates the model's superior performance in capturing the intricacies of human emotions.**

***Index Terms*—Fuzzy-deep neural network, multimodal data, multiscale feature, sentiment analysis.**

I. Introduction

In RECENT years, the field of sentiment analysis has witnessed significant advancements [1], fueled by the growing availability of multimodal data and the emergence of sophisticated deep learning techniques [2]. Sentiment analysis, also known as opinion mining, plays a pivotal role in understanding and interpreting human emotions expressed in textual, visual, and auditory content. This article enhances sentiment analysis accuracy and comprehensiveness by integrating multimodal data and multiscale features. We propose an innovative approach that combines multimodal representation and multiscale feature extraction using a fuzzy-deep neural network (Fuzzy-DNN) learning paradigm.

The proliferation of social media platforms, online reviews, and diverse forms of user-generated content has led to an exponential increase in the volume and variety of multimodal data expressing sentiments. Traditional sentiment analysis methods often rely solely on textual information, neglecting the valuable cues embedded in images, videos, and audio clips. Moreover, emotions are inherently complex and dynamic, requiring a nuanced understanding beyond single-scale analyses' limitations. The motivation for this research stems from the need to overcome these limitations, explore the untapped potential of combining multiple modalities, and consider the multiscale feature of emotional expressions.

Multimodal data encompasses various sources of information, including text, images, audio, and video [3]. Each modality contributes unique insights into the user's sentiments, which provides a holistic perspective on their emotional state. Text data is rich in explicit sentiment indicators such as sentiment-laden words and phrases. The model can capture explicit sentiments expressed through words by analyzing the linguistic content. Text directly represents the user's thoughts and feelings, making it a fundamental component of sentiment analysis. The proposed model uses text to capture these explicit sentiments, forming our sentiment analysis's basis. Audio data, particularly speech, contains prosodic features such as pitch, tone, and pace, which convey emotional nuances that text alone might miss. By incorporating audio, our model gains insights into the speaker's emotional state, enhancing its ability to interpret sentiments accurately. Image data, especially facial expressions and gestures, provide visual cues that are powerful indicators of emotions. Visual features such as smiles, frowns, and eye movements can significantly enhance sentiment detection. Visual cues are essential for understanding the emotional context in communication. Additionally, emotions are often manifested across different scales, ranging from subtle nuances to intense expressions. By incorporating multiscale features, the proposed approach aims to capture the richness of emotional experiences, offering a more nuanced and accurate sentiment analysis.

Manuscript received 12 March 2024; revised 16 May 2024 and 30 May 2024; accepted 22 June 2024. Date of publication 26 June 2024; date of current version 3 January 2025. This work was supported by the National Natural Science Foundation of China under Grant 62202247. Recommended by Associate Editor G. Srivastava. *(Corresponding author: Jianhui Lyu.)*

Xin Wang is with Northeastern University, Shenyang 110819, China (e-mail: dnsy_heinrich@neueet.com).

Jianhui Lyu and Qing Li are with Peng Cheng Laboratory, Shenzhen 518057, China (e-mail: lvjianhui2012@163.com; liq@pcl.ac.cn).

Byung-Gyu Kim is with Sookmyung Women's University, Seoul 04310, Republic of Korea (e-mail: bg.kim@sookmyung.ac.kr).

B. D. Parameshachari is with the Nitte Meenakshi Institute of Technology, Bangalore 560064, India (e-mail: paramesh@nmit.ac.in).

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/TFUZZ.2024.3419140

Fuzzy logic, inspired by human reasoning and decision-making processes, introduces a level of ambiguity into the analysis, allowing for the representation of partial truths and degrees of membership [4], [5]. In the context of sentiment analysis, where emotions often exhibit gradations and shades, fuzzy logic becomes a valuable tool for modeling the uncertainty inherent in human expressions. Fuzzy logic enables the Fuzzy-DNN model to handle linguistic terms [6], vague boundaries between sentiment categories, and the subtle transitions between different emotional states. Deep neural networks [7], on the other hand, excel at automatically learning hierarchical representations from complex data. The depth of these networks enables them to capture intricate patterns and dependencies within the input data. In the proposed Fuzzy-DNN framework, the deep neural network component serves as a feature extractor, hierarchically learning relevant features from multimodal and multiscale input data.

The integration of fuzzy logic principles with deep neural networks in the Fuzzy-DNN model is designed to harness the complementary strengths of both paradigms [8]. Fuzzy logic facilitates the modeling of uncertainty and imprecision, which allows the network to handle the inherent ambiguity in sentiment expressions. In our model, fuzzy sets are used to transform input features into degrees of membership, providing a nuanced representation of sentiment data. This approach allows for partial memberships, capturing the inherent uncertainty and ambiguity in emotional expressions. Our adoption of fuzzy sets in the model is motivated by their ability to handle uncertainty and ambiguity in sentiment data, reduce complexity in sentiment representation, and increase prediction accuracy. By providing a more flexible and nuanced approach to sentiment analysis, fuzzy sets enhance the model's overall performance and robustness. Meanwhile, the deep neural network component empowers the model to automatically learn complex representations and capture hierarchical dependencies, ensuring a robust and discriminative understanding of the underlying sentiment. One notable advantage of Fuzzy-DNN learning lies in its adaptive nature. The model can dynamically adjust its parameters and decision boundaries based on the contextual nuances of the input data, making it particularly well-suited for sentiment analysis tasks where emotions may be context-dependent. In addition, the fuzzy logic component enhances the interpretability of the model, providing insights into the degree of certainty associated with each sentiment prediction, a crucial aspect in understanding and trusting the model's decisions. The fusion of fuzzy logic principles with deep neural networks presents a powerful framework for handling uncertainty and imprecision inherent in sentiment analysis tasks. Fuzzy-DNN models excel in capturing the vagueness and fuzziness associated with human emotions, enabling a more flexible and adaptive representation of sentiment in multimodal data.

In this article, we combine the advantages of Multimodal and Multiscale features to develop an emotion analysis framework based on Fuzzy-Deep neural Network model research (MMFDN). Leveraging synergies between multimodal data and multiscale features, the integration of this fuzzy DNN learning paradigm aims to enhance the robustness of sentiment analysis models, thereby enabling them to better handle the complexity inherent in various forms of user-generated content. Our model integrates data from text, images, and audio to form a cohesive multimodal representation. The fusion of these modalities forms a unified representation input to the Fuzzy-DNN, ensuring a comprehensive sentiment analysis that considers the multifaceted nature of human expression. The multiscale feature extraction aims to capture nuances at various levels of granularity and enrich the representation of sentiments. This involves extracting features that represent both subtle nuances and prominent characteristics within the data.

We summarize the main contributions of this article as follows.

1) We introduce a novel sentiment analysis model that seamlessly integrates multimodal multiscale features from diverse modalities, including text, images, and audio, which creates a holistic representation of human emotion across different scales. Unlike traditional models, our approach utilizes fuzzy logic to handle uncertainty and a dual attention mechanism to dynamically focus on relevant features within and across modalities.
2) Our work incorporates fuzzy logic principles into the deep learning paradigm, presenting the fuzzy-deep neural network learning module. This addition enhances our model's adaptability to the inherent vagueness and uncertainty within sentiment expressions. This aspect distinguishes our approach from conventional deep learning models that do not explicitly address uncertainty in the same way.
3) We combine the dual attention mechanism module (DAM) to dynamically focus on crucial aspects within multimodal data. DAM optimizes attention allocation, refining feature extraction and fostering enhanced context-awareness in our model. This mechanism enhances the model's ability to extract meaningful information, thereby improving accuracy and robustness.
4) We comprehensively validate the proposed model on three distinct datasets, showcasing its superior performance across a spectrum of multimodal sentiment expressions.

The rest of this article is organized as follows. Section II provides a literature review, highlighting the existing approaches to sentiment analysis and the state-of-the-art techniques in multimodal and multiscale feature extraction. Section III details the proposed methodology, including the multimodal multiscale feature representation, fuzzy-deep neural network learning, and the dual attention mechanism. Section IV presents experimental results and discusses the findings. Finally, Section V concludes this article.

II. Related Work

Currently, numerous scholars both domestically and internationally have extensively contributed to the field of single-modal emotion analysis [9]. Considerable progress has been made in researching emotion analysis within text, speech, and image modalities, particularly with the notable integration of deep learning technology in recent years, significantly enhancing the precision of emotion recognition. However, in real-life scenarios, the expression of emotions is inherently diverse, and attempting to predict emotions from a singular perspective is akin to catching only a fleeting glimpse. Consequently, contemporary researchers increasingly advocate for emotion recognition grounded in multimodal data.

Sentiment analysis based on text data has gained widespread application in fields such as data mining and business recommendation, establishing itself as a prominent research area within natural language processing (NLP). In comparison to speech and image data, text information is rich in semantic content, offering a wealth of emotional features for sentiment analysis models. Historically, researchers treated emotion analysis as an unsupervised learning problem. They initially constructed dictionaries containing emotion words and their corresponding labels, such as WordNet-Affect [10] and BosonNLP [11]. Subsequently, the text to be identified was matched against these dictionaries to determine its classification category. As the field progressed, the introduction of various machine learning algorithms, including support vector machine (SVM) [12] and Bayesian methods, ushered in the era of supervised learning in text sentiment analysis.

In recent years, there has been a surge in the development of deep learning technologies, including CNN [13], LSTM [14], and Attention mechanisms [15]. The field of NLP has experienced rapid growth, leading to the continuous evolution of text sentiment analysis based on deep learning. For instance, Li et al. [16] utilized the attention mechanism and bidirectional LSTM model (Bi-LSTM) to focus the feature extraction network on the internal sequence relationships within the text. This approach effectively mitigated the issue of redundant text information, contributing to improved sentiment analysis. Furthermore, a prevalent and effective contemporary method involves leveraging word vectors to obtain semantic representations within sentences. This semantic information is then combined to derive comprehensive text features [17]. Notably, this approach finds widespread application across various multimodal sentiment analysis datasets.

At present, there are mainly three types of audio features widely used in the field of sentiment analysis, which are spectrum features, sound quality features, and prosodic features [18]. For example, Wang et al. [19] used a genetic algorithm and SVM as classifiers to classify prosodic features such as energy and fundamental frequency in audio. Their experiments proved that the optimization of parameters of SVMs by genetic algorithm could effectively improve the effect of sentiment classification. In addition, Eyben et al. [20] used LSTM and RNN networks to integrate prosodic features and semantic features in audio and achieved remarkable results. Since the extraction of audio features requires a certain understanding of acoustic knowledge, scholars often use audio feature extraction tools to extract audio features. At present, the mainstream collaborative speech analysis libraries include COVAREP [21] and openSMILE [22]. These toolkits can automatically extract key emotion-related features in audio, which greatly improves the efficiency of researchers. Some multimodal sentiment analysis datasets, such as UR-FUNNY [23] and MELD [24], use such audio feature extraction tools.

In the task of sentiment recognition, once speech features are extracted, an appropriate classification model is required for prediction and categorization. The prevailing classification models based on deep neural networks have demonstrated outstanding performance. For instance, Luo et al. [25] utilized CNN to extract features from the speech signal spectrum diagram. They input traditional acoustic features like speech spectrum centroid and MFCC into LSTM for deep feature extraction. Subsequently, they employed a Bi-LSTM algorithm with an attention mechanism for feature fusion and emotion classification, yielding remarkable results. In a similar vein, García-Ordás et al. [26] introduced a full-convolutional neural network as a classifier to predict and classify Maier spectrum features in audio, achieving commendable prediction outcomes.

Expression recognition based on deep learning has demonstrated remarkable results, involving a four-step process. First, the image data undergoes preprocessing operations such as scaling, cutting, alignment, and normalization. Subsequently, a face detection model is applied to identify the face in the image. Next, a feature extraction network extracts expression features, followed by the utilization of a classifier for sentiment classification. For instance, the PPDN network proposed by Zhao et al. [27] selects corresponding peak and nonpeak expression images from continuous video frames. The input feature extraction network then processes these images to extract features, using regularization to minimize the distance between the two sets of features. The cross-entropy function is ultimately employed to calculate the loss for the predicted results against the true value labels. However, it is worth noting that this model, relying on prefiltered expression images, exhibits reduced robustness in real-world scenes. To address this limitation, Ding et al. [28] introduced a novel training algorithm called FaceNet2ExpNet. This algorithm remodels neurons in the sentiment recognition model using information obtained from the face recognition model. By doing so, it adjusts the training of the sentiment recognition network. This innovative approach associates face feature extraction more closely with the classification model, resolving the issue of weak independent correlation between the two and thereby enhancing recognition effectiveness.

Although sentiment analysis based on single modalities such as text, speech, and images has made remarkable progress, there is a contradiction between the diversity of emotional expressions in real life and the limitations of single-modal data. Therefore, sentiment analysis based on multimodal data can better meet the needs of actual scenarios. At present, existing research on multimodal sentiment analysis mainly focuses on the two issues of representation learning of single-modal features and the fusion of multimodal features [29].

In terms of modal representation learning, Wang et al. [30] proposed the RAVEN network. They first extract single-modal features containing contextual information from speech and image data, and then dynamically convert the extracted nonlinguistic contextual features and text features to gain additional differentiation in single-modal representations. However, the difference obtained in this way is obtained by comparing text modalities, which neglects to capture the differences between nonlinguistic modalities, so the actual effect of this model is not

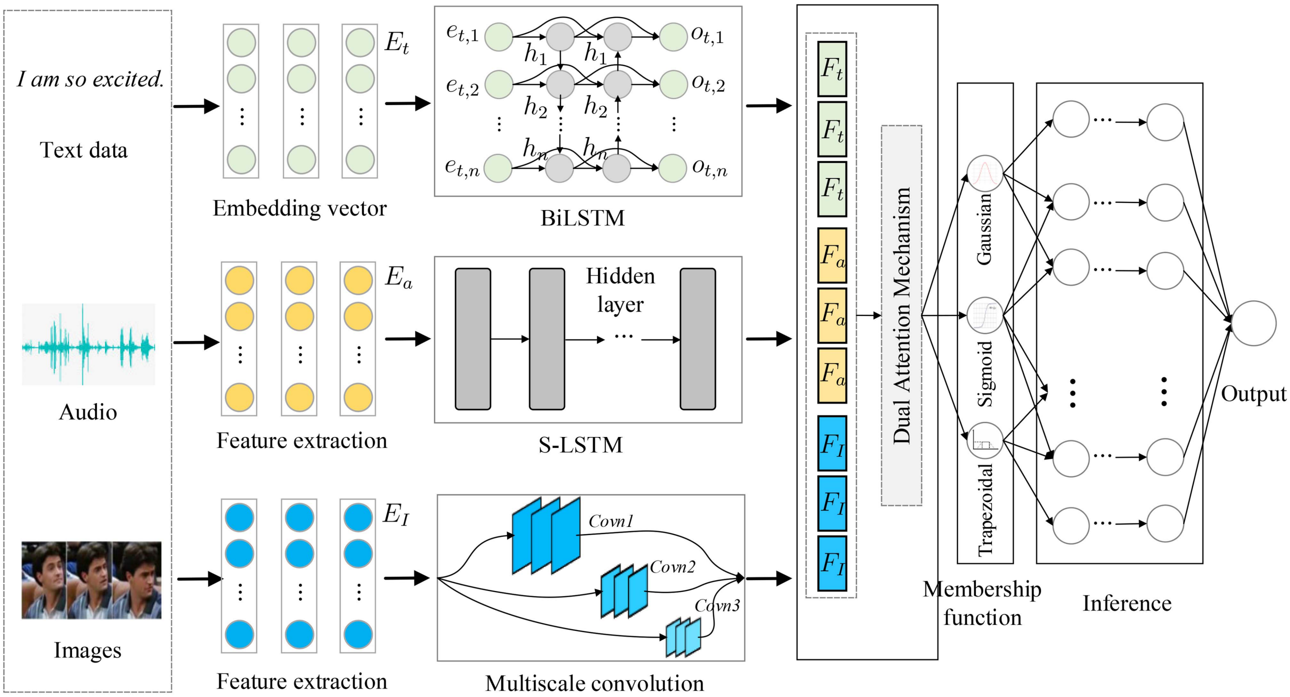

Fig. 1. Overall structure of the proposed model.

good. Verma et al. [31] respectively performed representation learning on common features between modalities and unique features within modalities of multimodal data. The obtained feature representations were both consistent and unique, and the sentiment recognition effect was significantly improved.

Feature layer fusion usually reorganizes features extracted from different modalities into a new feature representation through splicing and weighting operations. Most of the early feature layer fusion spliced features extracted from different modalities in the same dimension, then reduced the feature dimension, and finally input it into a classifier for sentiment analysis. For example, Pérez-Rosas et al. [32] extracted features from speech, text, and image data separately, then spliced the features, and finally used SVMs for sentiment classification. With the development of deep learning, single-modality feature extraction has changed from traditional manual feature extraction to automatic extraction using deep neural network training. The method of feature fusion has also evolved from simple splicing to deep fusion using various networks. For example, Zadeh et al. [33] designed a memory fusion network that can not only fuse feature representations between different modalities, but also model in the time dimension to obtain feature representations of contextual information. Tsai et al. [34] introduced a transformer-based network for multimodal feature extraction. It uses attention to better learn and combine features across modalities, improving the representation of target information.

Generally speaking, these methods have the following shortcomings. First, most current methods lack integration across modalities. In the context of sentiment analysis, a frequent issue is the insufficient integration of data from different modalities, leading to analyses that do not fully leverage the richness of multimodal information. Second, these methods cannot effectively extract multiscale features, particularly relevant to fields dealing with human language and emotion. That is, they cannot struggle to effectively handle the uncertainty and ambiguity inherent in natural language data.

III. Method

The complexity inherent in the analysis of sentiments expressed through multimodal data necessitates the development of advanced methodologies capable of capturing the nuances across various modalities and scales. In this section, we present a detailed account of the methodology employed in this study, focusing on the integration of fuzzy-DNN learning to explore and harness the potential of multimodal multiscale features for sentiment analysis, including the multimodal multiscale feature representation, fuzzy-deep neural network learning, and dual attention mechanism. The proposed model consists of several key components: convolutional layers for initial feature extraction, LSTM model for multiscale feature extraction, a dual attention mechanism for refining feature representation, and the fuzzy logic layer to handle uncertainty. The detailed architecture is shown in Fig. 1. Each convolutional layer is followed by a ReLU activation function. The overall structure of the proposed model is depicted in Fig. 1.

A. Multimodal Multiscale Feature Representation

The essence of our approach lies in its ability to seamlessly integrate information from different modalities–text, image, audio–into a cohesive representation that captures the multifaceted nature of human expression. Textual data undergoes tokenization and embedding, while images are processed through the convolutional neural networks for feature extraction. Simultaneously, audio data is transformed with the LSTM model. The fusion of these modalities forms a unified multimodal representation that serves as input to the Fuzzy-DNN model. Recognizing the multiscale nature of emotional expressions, our methodology incorporates multiscale feature extraction techniques. Features are extracted at varying levels of granularity, capturing both subtle nuances and prominent characteristics within the input data. This multiscale feature set is designed to enrich the representation of sentiments, providing the Fuzzy-DNN model with a comprehensive view of the emotional content present in the multimodal data.

For text data, the feature extraction layer uses the Bi-LSTM and the attention mechanism to perform deep semantic extraction of feature vectors. We use vectors to fuse character vectors and word vectors, respectively, to obtain fused features. The text will not change with changes in specific downstream tasks, and has distinct serialization characteristics. The Bi-LSTM model has a series network structure and is very suitable for processing serialized data. Therefore, in this article, we choose the Bi-LSTM model to process character features and word features. The Bi-LSTM model realizes context by splicing feature vectors with forward and reverse LSTM models. Effective utilization of semantic features. The calculation process of the LSTM model is as follows:

$$\begin{cases} i_t = \mathrm{f}(W_i \times [h_{t-1}, x_t] + b_\mathrm{i}), \\ f_t = \mathrm{f}(W_\mathrm{f} \times [h_{t-1}, x_t] + b_\mathrm{f}), \\ o_t = \mathrm{f}(W_\mathrm{o} \times [h_{t-1}, x_t] + b_\mathrm{o}), \\ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_\mathrm{c} \times [h_{t-1}, x_t] + b_\mathrm{c}), \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{1}$$

where x_t is the input vector at time t; i_t, f_t, and o_t represent the input gate, forget gate, and output gate at the current time, respectively; W_i, W_f, and W_o represent the weight matrices of the input gate, forget gate, and output gate, respectively; b_i, b_f, b_o represents the bias vector of the input gate, forget gate, and output gate, respectively; c_t represents the memory unit at the current moment; *t*–1 represents the moment before the current moment; W_c and b_c represent the weight matrix and bias vector of the current information, respectively; $\mathrm{f}(\cdot)$ and $\tanh(\cdot)$ are activation functions; h_t is the output vector at the current moment; $\odot$ is the Hadamard product; $\times$ represents matrix multiplication. So we can get the calculation process of the Bi-LSTM model as follows:

$$\begin{cases} \overrightarrow{h}_t = \mathrm{LSTM}(\overrightarrow{h}_{t-1}, x_t), \\ \overleftarrow{h}_t = \mathrm{LSTM}(\overleftarrow{h}_{t+1}, x_t), \\ o_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \end{cases} \tag{2}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ represent the feature vectors obtained by the forward and reverse LSTM models, respectively; o_t is the feature vector obtained by the Bi-LSTM model at the current time t; t+1 represents the previous time of the current time.

We set the initial state of the Bi-LSTM model to 0. Then, the character vector set E_c and the word vector set E_w are input into the Bi-LSTM model, respectively, to obtain the character feature vector set $o_c = \{o_{c,1}, o_{c,2}, \ldots, o_{c,n}\}$, the word feature vector set $o_w = \{o_{w,1}, o_{w,2}, \ldots, o_{w,n}\}$, and the state S_c and S_w stored by the Bi-LSTM model. The calculation is as follows:

$$\begin{cases} o_{\mathtt{c},i} = \mathrm{BiLSTM}(e_{\mathtt{c},i}), 1 \leq i \leq n, \\ o_{\mathtt{w},j} = \mathrm{BiLSTM}(e_{\mathtt{w},j}), 1 \leq j \leq m \end{cases} \tag{3}$$

where $e_{\mathtt{c},i}$ and $e_{\mathtt{w},j}$ represent the vector in E_c and E_w, respectively.

The audio modality, a rich source of emotional cues, is processed using a unidirectional Long LSTM (S-LSTM) model to capture temporal dependencies and patterns inherent in speech. LSTMs, a variant of recurrent neural networks, are well-suited for sequential data processing due to their ability to selectively retain information over extended time intervals. The unidirectional architecture allows the LSTM to process the input sequence in a forward temporal direction, capturing the temporal dynamics of the audio signal. This approach leverages the ability of unidirectional LSTMs to effectively model the temporal dynamics of speech, providing a compact and informative representation of the audio modality. The extracted features contribute to the holistic multimodal representation, enabling the Fuzzy-Deep Neural Network model to discern and analyze sentiment across diverse modalities and scales.

Specifically, referring to the practice of Zadeh et al. [35], we use the S-LSTM network to model speech features in the time dimension and obtain their context representation. Its form is shown in the following formula:

$$F_a = S\text{-}LSTM(E_a; \theta_a) \in R^{d_a} \tag{4}$$

where E_a means the audio feature, θ_a represents its hidden layer parameters, d_a represents the audio data dimension, and F_a is the hidden layer output sequence of the voice modality, which represents the context feature representation of a single modality.

When extracting features from images, only extracting a single feature is not well suited to the image. Zhang et al. [36] decomposed medical images into multiple scale layers and can extract different visual features from different scale layers. Because image analysis needs to consider multiple aspects of information, this information is at different scale levels. Lin et al. [37] proposed Dual Swin Transformer UNet (DS-TransUNet) to extract coarse-grained feature representations and fine-grained feature representations of different semantic scales for image segmentation tasks. Kong et al. [38] used four different scaled histopathological images to generate four semantic feature maps of different sizes, so that the model has strong generalization ability from tissue types to cell types. Due to the particularity of images, when extracting image features and classifying them, researchers usually use a multiscale approach to make judgments, which helps determine whether there are obvious feature abnormal areas.

Inspired by the multiscale feature extraction method in image classification, in this article, we designed a multiscale feature extraction module. The structure diagram is shown in Fig. 1. This module uses a CNN-based multiscale feature extraction method to extract features on small-scale $F_{I,s}$, medium-scale $F_{I,m}$, and large-scale $F_{I,l}$, respectively. We exploit the encoder-decoder structure and extract image features to reduce feature scale through multiple convolution and pooling layers. In the decoder, we perform deconvolution and upsampling operations while extracting features that require scale. Furthermore, we keep the dimensions of the three granularity blocks all the same size during block embedding to reduce the fusion cost and thereby speed up the computation time. We utilize the skip connections to help gradients flow through the entire network and preserve the feature map, especially the detailed information in fine granularity.

Then, the multiscale features $F_{I,s}$, $F_{I,m}$, and $F_{I,l}$ are remolded into a series of flat 2-D feature blocks as follows:

$$\begin{aligned} &F_{I_p,s} \in \mathbf{R}^{\frac{H}{2r}\times\frac{W}{2r}\times C_{\mathrm{s}}}, \\ &F_{I_p,m} \in \mathbf{R}^{\frac{H}{r}\times\frac{W}{r}\times C_{\mathrm{m}}}, \\ &F_{I_p,l} \in \mathbf{R}^{H\times W\times C_{\mathrm{l}}}, \end{aligned} \tag{5}$$

where (*H*, *W*) represents the resolution of the original feature, *C* represents the number of channels. We enable different dimensions to be matched at different scales, which can significantly reduce computational complexity when fused with fuzzy features. We add a learnable embedding to each sequence of embedded feature blocks. The multiscale feature extraction module also uses the position embedding strategy when processing image data, i.e., position information is embedded into feature blocks to retain the accurate position of each pixel.

B. Fuzzy-Deep Neural Network Learning

We choose fuzzy set theory to solve the problem of information redundancy in multiscale feature extraction, because fuzzy set theory regards the pixel value of the image as an element and describes the membership degree of the element to the fuzzy concept through the membership function in image processing. The advantage of this method is that it can capture the uncertainty and fuzziness in the image. By setting different membership functions and critical values, the general characteristics of malignant cells can be accurately extracted and more accurate information can be provided for image classification.

For a given image I, first, we convert *I* into a grayscale image and normalize it to the range [0, 1]. When extracting the fuzzy features of the image, each pixel point x is regarded as a fuzzy set, and each image passes through a different membership function. In this article, we extract three fuzzy features f_μ, f_δ, and f_τ, and the obtained fuzzy universal feature set is expressed as $\{f_\mu, f_\delta, f_\tau\}$. The definition of this fuzzy set depends on the membership function, which describes the degree of membership of each pixel to a certain fuzzy concept. Finally, the fuzzy general feature f_z is obtained through the fuzzy operation.

During the model learning process, the membership function can help the model better understand the meaning of each pixel, including its possible category and degree of belonging. This multiangle and multilevel feature expression method can provide more information, help the model capture richer and more complex features, and also improve the model's robustness to noise and uncertainty. In order to extract multiple features of the image, we use multiple membership functions, each membership function corresponds to a specific feature description method. We select Gaussian function, Sigmoid function, and Trapezoidal function as membership functions to fuzzify the image, respectively, so that fuzzy general features can more effectively guide the model to learn key features. These features are used to construct fuzzy sets, and the common features of the image are extracted through fuzzy set intersection operations. The Gaussian function is chosen to define the first membership function because the Gaussian function has good smoothness and symmetry. It can calculate the gray value of each pixel point with the membership function to obtain the membership degree of each pixel point. Its calculation process is as follows:

$$F_\mu(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\boldsymbol{\mu})^2}{2\sigma^2}\right) \tag{6}$$

where μ represents the mean of the Gaussian distribution, σ represents the standard deviation, and *x* represents the gray value of the pixel. The Sigmoid membership function is defined as follows:

$$F_\delta(x) = \frac{1}{1+\exp(-\alpha(B_x-\beta))} \tag{7}$$

where B_x represents the brightness value of pixel point x, α, and β represent the parameters of the function, which are used to adjust the shape and position of the function. Finally, the Trapezoidal membership function is defined as follows:

$$F_\tau(x) = \begin{cases} 0, & x \le a \quad or \quad x \ge d \\ \frac{x-a}{b-a}, & a < x \le b \\ 1, & b < x \le c \\ \frac{d-x}{d-c}, & c < x < d \end{cases} \tag{8}$$

where a, b, c, and d represent the rising slope, falling slope, falling inflection point, and endpoint critical value, respectively. The left and right ends of the function are controlled by the rising slope a and the falling slope b, respectively, and the middle part is 1. When the gray value of a pixel is less than a or greater than d, the membership degree is 0, i.e., it does not belong to the fuzzy attribute at all. When the value is between a and b, the degree of membership gradually increases. When the value reaches b, the membership degree reaches 1, which means it completely belongs to the fuzzy attribute. When the value is between b and c, the membership degree is always 1, i.e., it completely belongs to the fuzzy attribute. When the value is between c and d, the degree of membership gradually decreases. When the value reaches d, the membership degree is 0, i.e., it does not belong to the fuzzy attribute at all. Through fuzzy set theory, the features extracted by various membership functions are effectively integrated to form a more comprehensive and accurate general feature. This general feature can reflect more information and better guide the learning of the model. For the Gaussian function, we choose the mean μ and standard deviation σ to be 0 and 1, respectively. For Sigmoid function, the parameter values of α and β are 1 and 0, respectively. For the Trapezoidal function, the parameter values of parameters a, b, c, and d are 0.1, 0.3, 0.6, and 0.9, respectively.

Let F be the universe of discourse for textual sentiment, where $F = \{f \in \mathbb{R} \mid -1 \le f \le 1\}$. Here, –1 represents the most negative sentiment, 0 represents a neutral sentiment, and 1 represents the most positive sentiment. In order to obtain fuzzy universal features of images, we also introduce a fuzzy weighting strategy to fuse uncertain data, so as to integrate these three features and form a more comprehensive feature description. Specifically, a weight is set for each fuzzy feature, namely w_μ, w_δ, and w_τ, which reflect the importance of each feature to the overall description and satisfy $w_\mu + w_\delta + w_\tau = 1$. In addition, we introduce an offset b_z to adjust the baseline level of fuzzy feature fusion to improve the flexibility of the model. Through the above variables and parameters, the fuzzy general features are obtained as follows:

$$F_{fz} = w_\mu f_\mu + w_\delta f_\delta + w_\tau f_\tau + b_z. \tag{9}$$

In fuzzy fusion, the membership degree of a data point determines its contribution to the average value. This method can effectively integrate the features extracted by multiple membership functions to obtain a more comprehensive image representation. In this way, rich fuzzy features can be extracted from the image and can be used for subsequent model learning.

C. Dual Attention Mechanism

The attention mechanism in deep learning mainly draws on the visual attention mechanism that allows humans to quickly filter out important information from a large amount of information. For example, we can distinguish the primary and secondary content in an image and only pay attention to and process a small number of words to be read when reading an article. This idea was first proposed in the field of computer vision. By learning the weight distribution of image positions and features, the features are weighted, so that deep learning tasks can be prioritized and work efficiency improved. For example, the literature [39] used the attention mechanism to learn the position of the image to be processed. On this basis, authors in [40] proposed a convolutional attention module that combines the spatial domain and the channel domain. This is a lightweight and efficient attention mechanism, which sequentially deduces the attention map along two independent dimensions (channel and spatial) and multiplies it with the input feature map for adaptive feature optimization, so that the network can achieve better accuracy and effectiveness.

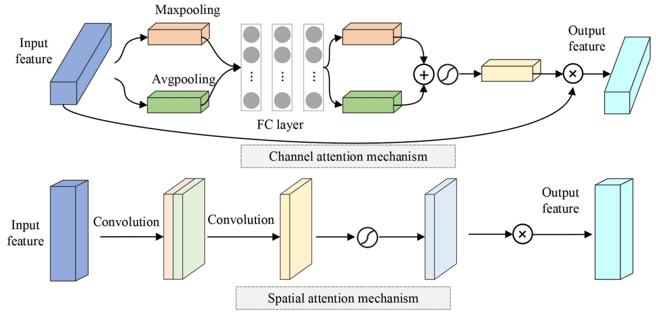

Fig. 2. Detailed structure of the dual attention mechanism.

With this advantage, we design a dual attention mechanism to enable the model to better extract features and complete subsequent tasks. The dual attention mechanism mainly includes the channel attention mechanism and the spatial attention mechanism, as shown in Fig. 2. The channel-wise attention mechanism mainly acts on different convolution channels of the input feature map. The contribution of features to key information on each channel is different, which is reflected by adding a weight to each channel. Its principle can be understood from the perspective of signal system analysis. Assuming that an image generates new channels after passing through different convolution kernels, the image features of each channel are equivalent to its components on different convolution kernel functions. This is similar to time-frequency transformation, which adds weights to the signal component on each channel to represent the correlation between the channel and the key information. The main work of the channel attention module can be described as three steps: 1) compression; 2) excitation; and 3) scale. The first step of the compression operation is to perform global average pooling and global maximum pooling operations on the global spatial features of each channel. Assume that the size of the input feature map F_{map} is $H \times W \times C$. After the compression operation, two feature maps of $1 \times 1 \times C$ are obtained, which realizes the function of aggregating spatial information. The next step is to pass these two feature maps through a two-layer fully connected neural network to perform element-wise summation operations and function activations, which can be expressed as follows:

$$C_{\text{att}} = f_{\text{sig}}\left(W_{a_1} \cdot \gamma(W_{a_0} F_{\text{avg}}) + W_{a_1} \cdot \gamma(W_{a_0} F_{\max})\right) \tag{10}$$

where f_{sig} and $\gamma(\cdot)$ are the Sigmoid and ReLU activate functions, respectively, F_{avg} and $F_{\max}$ are the feature vectors after the average pooling layer and the maximum pooling layer, respectively, W_{a_0} and W_{a_1} are the weights. The last step is to reweight the features and multiply the learned attention map with the output F of the feature fusion module to obtain the features containing the channel attention representation. The specific formula is as follows:

$$F_{\text{att}_c} = C_{\text{att}} \cdot F \tag{11}$$

where F is the concatenation of text, speech, and image features as follows:

$$F = concat(F_t, F_a, F_I) \tag{12}$$

where $concat(\cdot)$ represents the concatenation operation.

The spatial-wise attention mechanism focuses on finding position information in the input features that are highly relevant to the task target. After the network obtains the attention mask through learning, it uses weighting to highlight the important feature spatial locations in the feature map. First, we perform channel-based global maximum pooling and global average pooling on the feature map F_{att_c} weighted by channel attention to obtain F'_{avg} and $F'_{\max}$. Then, we concatenate the two to obtain $H \times W \times 2$ channel features and use a convolution operation to compress the number of channels to 1. Finally, function activation is used to obtain the spatial attention weight, which is multiplied by the input feature F_{att_c} to obtain a feature map with spatial attention representation. The specific formula is as follows:

$$V_{\text{att}} = f_{\text{sig}}\left(covn(concat(F'_{\text{avg}}, F'_{\max}))\right) \tag{13}$$

$$F_{\text{att}_v} = V_{\text{att}} \cdot F_{\text{att}_c} \tag{14}$$

Algorithm 1: The Algorithm of the Multimodal Sentiment Analysis.

```
Require: Multimodal data: Text $d_t$, Audio $d_a$, Image $d_I$
Ensure: Sentiment prediction $y_s$
for each data sample in $\{d_T, d_A, d_I\}$ do
    $E_t \leftarrow FeatureExtractText(d_t)$
    $E_a \leftarrow FeatureExtractAudio(d_a)$
    $E_I \leftarrow FeatureExtractImage(d_a)$
    $F_t \leftarrow BiLSTM(E_t)$
    $F_a \leftarrow S-LSTM(E_a)$
    $F_I \leftarrow MultiscaleConv(E_I)$
end for
for each feature $f$ in $\{F_t, F_a, F_I\}$ do
    $F_{att} \leftarrow Attention(f)$
    $F_{f_z} \leftarrow Fuzzify(F_{att_c})$
end for
$F_{fused} \leftarrow FuseFeatures(F_{f_z})$
$y_s \leftarrow$ Classify$(F_{fused})$
Return $y_s$
```

where $covn(\cdot)$ means the convolution operation block.

We consider introducing a dual attention mechanism module based on multiscale feature fusion, so that it can pay more attention to feature representations with strong feedback capabilities and discriminability. During the learning process, the network can adaptively adjust the attention weights of features of different scales according to the importance of the features, suppressing the interference of nonkey features with lower weights, which is more conducive to the emotional analysis of images. The detailed algorithmic pseudocode of the method is presented in Algorithm 1.

IV. Experiments

In this section, we present a comprehensive overview of the experimental methodology employed to evaluate the proposed multimodal sentiment analysis framework, integrating fuzzy-deep neural network learning with a focus on multimodal and multiscale features. The experiments are designed to assess the model's efficacy in capturing the intricacies of sentiment expressed through diverse modalities, emphasizing textual, visual, and auditory inputs. The following subsections detail the key components of our experimental setup, including the dataset description, evaluation metrics, experimental environment, performance analysis, and crucially, the conduct of ablation experiments to unravel the individual contributions of different components to the overall model performance.

A. Experimental Setup

1) Configuration: The computational infrastructure and software environment play a pivotal role in ensuring the reproducibility and scalability of our experiments. This subsection provides a detailed description of the hardware and software configurations used, including the specifications of the machines, the deep learning frameworks leveraged, and any additional libraries employed for efficient implementation. Our comprehensive experiments unfolded in a thoughtfully crafted environment, strategically blending cutting-edge hardware and software configurations to propel the research into multimodal sentiment analysis. At the heart of our computational prowess are the NVIDIA Tesla V100 GPUs, renowned for their parallel processing capabilities, with each GPU boasting an impressive 32 GB of memory. The hardware setup included an Intel Core i9-10900 K processor with 10 cores at a base frequency of 3.7 GHz. The system is equipped with 32 GB DDR4 RAM at 3200 MHz, ensuring efficient handling of large datasets and complex algorithms. This high-performance computing cluster provided the robust foundation necessary for accelerated training of our intricate deep neural network models, prominently the Fuzzy-DNN architecture. In conjunction with this potent hardware infrastructure, PyTorch 1.7 served as our deep learning framework, chosen for its dynamic computational graph and extensive toolkit. Complementing PyTorch, CUDA, and CuDNN is integral to harnessing the parallel computing power of our NVIDIA GPUs, resulting in a significant reduction in model training times.

2) Metrics: Quantitative assessment of the proposed multimodal sentiment analysis framework involves the application of key performance metrics, each offering unique insights into the model's efficacy. The selected metrics encompass accuracy (Acc), precision (Prc), recall (Rec), and F1 score, collectively forming a comprehensive evaluation framework.

Accuracy serves as a fundamental measure of the model's overall correctness in predicting sentiments across all classes. It is defined as the ratio of correctly predicted instances to the total number of instances in the dataset.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP denotes True Positives, TN True Negatives, FP False Positives, and FN False Negatives.

Precision gauges the model's ability to correctly identify positive instances among those predicted as positive. It is calculated as the ratio of True Positives to the sum of True Positives and False Positives.

$$\text{Prc} = \frac{TP}{TP + FP}.$$

Recall, also known as Sensitivity or True Positive Rate, measures the model's capability to capture all positive instances in the dataset. It is computed as the ratio of True Positives to the sum of True Positives and False Negatives.

$$\text{Rec} = \frac{TP}{TP + FN}.$$

The F1 score represents the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is particularly useful in scenarios where there is an imbalance between positive and negative instances.

$$\text{F1} = \frac{2 \cdot \text{Prc} \cdot \text{Rec}}{\text{Prc} + \text{Rec}}.$$

TABLE I
DETAILS OF THE SAMPLE SIZE OF THE MOSI DATASET

No.	Type	Grade	Number	Description
1	Negative	-3	185	Very negative
2	Negative	-2	440	Negative
3	Negative	-1	398	More negative
4	Middle	0	96	Neutral state
5	Positive	1	361	More positive
6	Positive	2	482	Positive
7	Positive	3	237	Very positive
All	-	-	2199	-

TABLE II
DETAILS OF THE SAMPLE SIZE OF THE MOSEI DATASET

No.	Type	Grade	Number	Description
1	Negative	-3	816	Very negative
2	Negative	-2	2231	Negative
3	Negative	-1	3547	More negative
4	Middle	0	4998	Neutral state
5	Positive	1	7429	More positive
6	Positive	2	3170	Positive
7	Positive	3	665	Very positive
All	-	-	22856	-

TABLE III
DETAILS OF THE SAMPLE SIZE OF THE SIMS DATASET

No.	Type	Grade	Number	Description
1	Negative	(-1,0)	754	Negative
2	Negative	(-1,0)	484	More negative
3	Middle	0	345	Neutral state
4	Positive	(0,1)	346	More positive
5	Positive	(0,1)	352	Positive
All	-	-	2281	-

In these equations, TP refers to True Positives (correctly predicted positive instances), TN to True Negatives (correctly predicted negative instances), FP to False Positives (incorrectly predicted positive instances), and FN to False Negatives (incorrectly predicted negative instances). These metrics collectively offer a nuanced evaluation of the model's performance, reflecting its ability to accurately discern and classify sentiments across diverse multimodal inputs.

3) Database: In order to comprehensively verify the performance of the model proposed in this article, we used three different datasets to verify and analyze the performance of the model in the laboratory, including the Multimodal Corpus of Sentiment Intensity (MOSI) dataset [41], the Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset [42], and the Chinese Single and Multimodal Sentiment (SIMS) dataset [43]. MOSI dataset is one of the mainstream datasets for multimodal sentiment analysis, which promotes the development of multimodal sentiment analysis. The MOSI dataset consists of 93 Youtube movie review videos from 89 speakers, 48 men, and 41 women, aged between 20 and 30. The dataset segments the film review videos in one-sentence units, and takes the segmented short videos as the raw data. Then, the short video is processed by feature extraction tools such as speech, text, and image, and then the text, speech, and image data samples are obtained. The sample size of the three modes is 2199. The MOSI dataset uses manual labeling to score the sentiment of each piece of data on a scale of –3 to 3. A lower score indicates a more negative emotion, a higher score indicates a more positive emotion, and 0 indicates neutrality. The data distribution of different categories in the MOSI dataset is shown in Table I. MOSEI dataset is a multimodal sentiment analysis dataset, an upgraded version of MOSI dataset, which has three modes: 1) text; 2) speech; and 3) image. The dataset consisted of 23 452 manually tagged video clips from more than 1000 Youtube narrators covering 250 speaking topics over a total of 65 hours. In this dataset, each sample is labeled with seven categories of emotion. According to the strength of emotion, the label value is between –3 and 3, where 0 indicates neutrality. The dataset contains 16 326 training samples, 1871 verification samples, and 4659 test samples. The dataset description is shown in Table II. The SIMS dataset]is a multimodal sentiment analysis dataset of Chinese corpus, which contains text, speech, and image modalities. This dataset not only uniformly annotates datasets of different modalities, but also annotates each single modality separately. SIMS contains 60 video clips with 474 narrators selected from Chinese movies, TV series, and reality shows. This dataset segments videos according to sentences and annotates emotions according to five categories. Annotators score the emotional intensity of video content on a scale from –1 to 1, with 0 indicating neutral. There are 2281 samples in the SIMS dataset, including 1368 training sets, 456 validation sets, and 457 test sets. The details of each category in the data sample are shown in Table III.

We preprocess the data before inputting it into the network model as follows. For text data, preprocessing steps for text data include tokenization, stemming, and vectorization using term frequency–inverse document frequency. We segmented the text into words or phrases to simplify further processing. Then, the processed text was converted into numerical values using techniques such as word embeddings to prepare it for input into the neural network. For audio data, we applied filtering techniques to remove background noise and enhance the clarity of the speech signals. Then, key features such as Mel-frequency cepstral coefficients (MFCCs) were extracted to represent the speech data efficiently. The features were normalized to have zero mean and unit variance to facilitate model training. For image data, all images were resized to a standard dimension 224 × 224 to ensure uniformity across the dataset. Pixel values were normalized to the range [0, 1] to improve model convergence during training. Each dataset is divided into training, validation, and test sets, with proportions of 70%, 15%, and 15%, respectively. The training procedure involves using the Adam optimizer with a learning rate of 0.0001, and the model is trained for 100 epochs with early stopping based on validation loss.

B. Performance Comparison

In order to verify the performance of the multimodal and multiscale sentiment analysis method based on fuzzy deep]neural network representation learning proposed in this article, we conducted comparative experiments using the MOSI, MOSEI, and SIMS datasets, respectively. We have reproduced several currently mainstream and effective multimodal sentiment analysis models, including Multimodal Factorization Model (MFM),

TABLE IV
COMPARISON RESULTS WITH CURRENTLY MAINSTREAM MULTIMODAL SENTIMENT ANALYSIS METHODS BASED ON THE MOSI DATASET

References	Method	Types	Acc (%)	Prc (%)	Rec (%)	F1 (%)
Tsai et al. [44]	MFM	7	37.45	41.82	46.11	43.86
Sun et al. [45]	ICCN	7	42.67	43.28	49.51	46.19
Hazarika [46]	MISA	7	41.65	44.71	44.69	44.70
Rahman et al. [47]	MAG-BERT	7	46.72	50.33	53.98	52.09
Yu et al. [48]	Self-MM	7	49.83	48.72	49.57	49.14
Han et al. [49]	MMIM	7	56.63	52.21	56.72	54.37
Jochen et al. [50]	SiEBERT	7	58.69	**58.28**	57.32	57.80
Wu et al. [51]	KDGN	7	57.46	57.18	57.07	57.12
Ours	MMFDN	7	**59.74**	57.17	**60.78**	**58.92**

Interaction Canonical Correlation Network (ICCN), Modality Invariant and Specific Representations (MISA), Multimodal Adaption Gate for BERT (MAG-BERT), Self-supervised Multi-Task Learning (Self-MM), MultiModal InfoMax (MMIM), SiEBERT, and KDGN. A brief introduction to these baseline methods follows. The MFM model connects the inference network and the generation network by establishing intermediate modal factors and uses the reconstruction loss function and the discrimination loss function to promote the fusion process. The ICCN model introduces canonical correlation analysis loss between different modes to improve the fusion effect of the model. MISA projects multimodalities into modality-invariant and modality-specific spaces, respectively, captures modality-shared and modality-unique features, respectively, and uses these overall features to fuse and predict results. The MAG-BERT method designs a multimodal modal alignment threshold network and embeds it into the BERT model to optimize the overall fusion effect. The Self-MM model assigns each modality a task that can automatically generate the corresponding modal label and adjusts the gradient of backpropagation through multitask learning. The MMIM model solves the mutual information between each modality and the multimodal fusion results and the mutual information between each modality and uses multitask learning to maximize the mutual information to improve the fusion effect. SiEBERT serves as an empirical framework that measures the tradeoffs across various research questions, data traits, and analytical capabilities, facilitating the selection of methods based on the specific context of their application. KDGN improves the model performance based on the dependency graph incorporating domain knowledge.

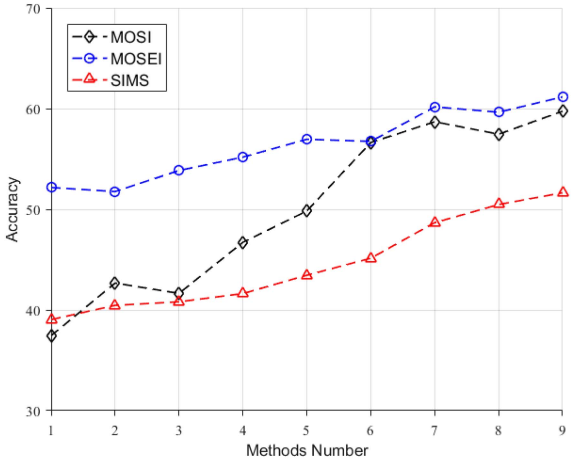

Fig. 3. Experimental comparison accuracy results of compared methods based on three datasets. The numbers 1 to 9 on the horizontal axis represent methods MFM, ICCN, MISA, MAG-BERT, Self-MM, MMIM, SiEBERT, KDGN, and MMFDN, respectively.

Table IV lists the experimental effects of each baseline method and the method proposed in this article on the MOSI dataset. It can be seen from the table that MFN, MISA, ICNN, and other methods perform poorly, because these methods can only extract the interaction information between multiple modes, and lack effective characterization of the internal information of a single mode. In contrast, Self-MM and MMIM can better learn the emotional information within and between modes, so they have better performance. The representation learning method based on multimodal and multiscale features proposed in this article can not only effectively represent a single mode, but also fully integrate features between multiple modes. Its classification accuracy is 13% (59.74% versus 46.72%), 10% (59.74% versus 49.83%), and 3% (59.74% versus 56.63%) higher than that of MAG-BERT, Self-MM, and MMIM with better performance, respectively. Table V lists the experimental results of each baseline method on the MOSEI dataset. Compared with the MOSI dataset, the MOSEI dataset has more data and a greater amount of information that can be learned. Therefore, each indicator on the MOSEI dataset is better than the MOSI dataset. Our method still performs well compared with the baseline method, with classification accuracy reaching 61.16% and F1-score reaching 60.36%, respectively. Since MOSI and MOSEI are both English corpus datasets, in order to verify the performance of our method in Chinese corpus datasets, we utilize the SIMS dataset to compare our method with the model in the above experiment. The experimental results are shown in Table VI. The experimental results show that in the SIMS dataset, the performance gap of different methods is not obvious in the MOSI and MOSEI datasets, because the emotion category and distribution of the SIMS dataset are quite different from other datasets, and the methods are not sensitive to the SIMS dataset. Compared with other methods, our proposed method MMFDN still has improved performance and achieved the best classification accuracy. In order to compare the performance trends of each method more clearly, we plotted the accuracy and F1 score indicators of each method on the three datasets, as shown in Figs. 3 and 4.

TABLE V
COMPARISON RESULTS WITH CURRENTLY MAINSTREAM MULTIMODAL SENTIMENT ANALYSIS METHODS BASED ON THE MOSEI DATASET

References	Method	Types	Acc (%)	Prc (%)	Rec (%)	F1 (%)
Tsai et al. [44]	MFM	7	52.16	49.94	47.68	48.78
Sun et al. [45]	ICCN	7	51.76	51.17	49.27	50.20
Hazarika [46]	MISA	7	53.85	53.24	55.92	54.55
Rahman et al. [47]	MAG-BERT	7	55.18	52.94	53.47	53.20
Yu et al. [48]	Self-MM	7	56.95	57.67	58.16	57.91
Han et al. [49]	MMIM	7	56.74	55.32	59.65	57.40
Jochen et al. [50]	SiEBERT	7	60.17	58.51	61.38	59.91
Wu et al. [51]	KDGN	7	59.64	**58.81**	60.72	59.75
Ours	MMFDN	7	**61.16**	58.39	**62.46**	**60.36**

TABLE VI
COMPARISON RESULTS WITH CURRENTLY MAINSTREAM MULTIMODAL SENTIMENT ANALYSIS METHODS BASED ON THE SIMS DATASET

References	Method	Types	Acc (%)	Prc (%)	Rec (%)	F1 (%)
Tsai et al. [44]	MFM	5	39.04	37.86	40.64	39.20
Sun et al. [45]	ICCN	5	40.46	39.47	41.16	40.30
Hazarika [46]	MISA	5	40.81	38.14	39.47	38.79
Rahman et al. [47]	MAG-BERT	5	41.63	42.68	42.24	42.46
Yu et al. [48]	Self-MM	5	43.45	45.46	43.63	44.53
Han et al. [49]	MMIM	5	45.14	43.17	46.78	44.90
Jochen et al. [50]	SiEBERT	5	48.67	49.51	**51.38**	50.43
Wu et al. [51]	KDGN	5	50.47	50.76	49.83	50.29
Ours	MMFDN	5	**51.64**	**52.62**	50.42	**51.50**

The bold fonts are the best results for the corresponding metrics to highlight the performance benefits of our model.

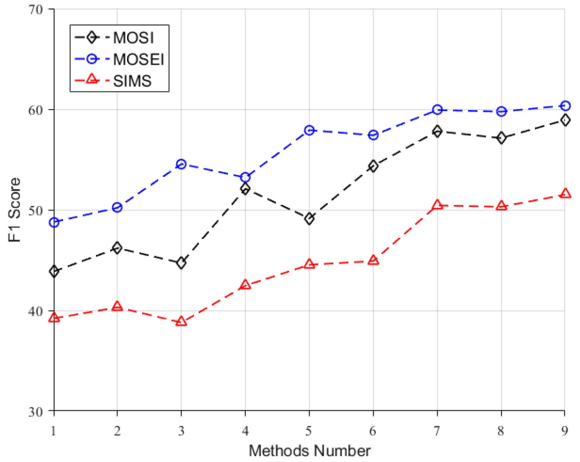

Fig. 4. Experimental comparison F1 score results of compared methods based on three datasets. The numbers 1 to 9 on the horizontal axis represent methods MFM, ICCN, MISA, MAG-BERT, Self-MM, MMIM, SiEBERT, KDGN, and MMFDN, respectively.

In our study, we compared the performance of our proposed MMFDN model with several state-of-the-art sentiment analysis models across three public datasets. The models used for comparison include MFM, ICCN, MISA, MAG-BERT, Self-MM, MMIM, SiEBERT, and KDGN. Our model consistently outperformed the compared models in terms of accuracy across all datasets. For instance, as shown in Tables V, the MMFDN achieved an accuracy of 61.16% on dataset MOSEI, compared to 60.17% for SiEBERT, 59.64% for KDGN, and 56.74% for MMIM. This improvement in accuracy can be attributed to the effective integration of multimodal data and the robustness of the fuzzy logic layer in handling uncertainty and imprecision in sentiment expressions. The precision and recall metrics indicate the model's ability to correctly identify positive and negative sentiments, respectively. Our model demonstrated superior recall and comparable Precision, particularly in cases with ambiguous or mixed sentiments. For example, in Table IV, our method shows a precision of 57.17% and a recall of 60.78% on dataset MOSI. This performance is enhanced by the dual attention mechanism, which allows the model to focus on the most relevant features within and across modalities. The F1-score, which is the harmonic mean of precision and recall, further highlights the balanced performance of our model. Our model achieves the best results on three datasets. This indicates that our model not only identifies sentiments accurately, but also maintains a good balance between precision and recall.

To delve deeper into image analysis and assess model performance more comprehensively, we conducted a statistical analysis. Fig. 3 presents the results of a statistical comparison between the proposed model and other models evaluated on the SIMS dataset. In the figure, A to H represent MFM, ICCN, MISA, MAG-BERT, Self-MM, MMIM, SiEBERT, and MMFDN methods, respectively. The white grid marks statistical equivalence between the method of the row and that of the column. Conversely, a grid shaded gray/black denotes the row method's statistical superiority/inferiority to the column method, respectively. We can see that the performance of the method MMFDN proposed in this article exceeds that of other methods in most indicators, which reveals the effectiveness of the method in sentiment analysis.

C. Importance of Multimodal and Multiscale Features

In this section, to quantify the distinct contributions of multimodal multiscale features in our proposed sentiment

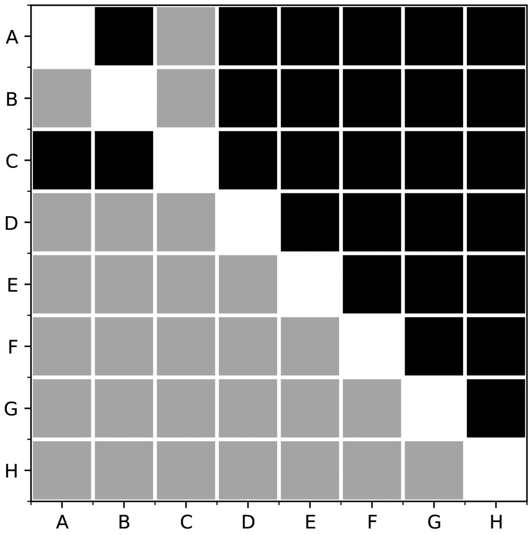

Fig. 5. Statistical analysis results based on the SIMS dataset.

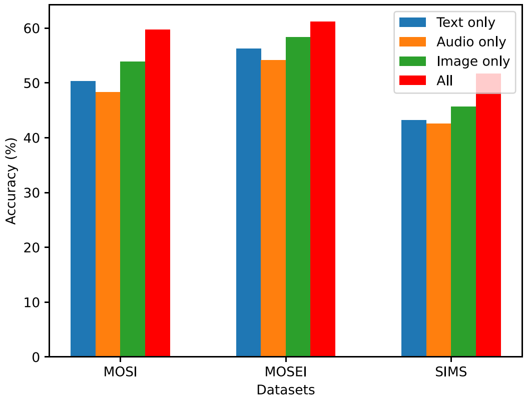

Fig. 6. Experimental comparison results of feature extraction in different modalities based on three datasets.

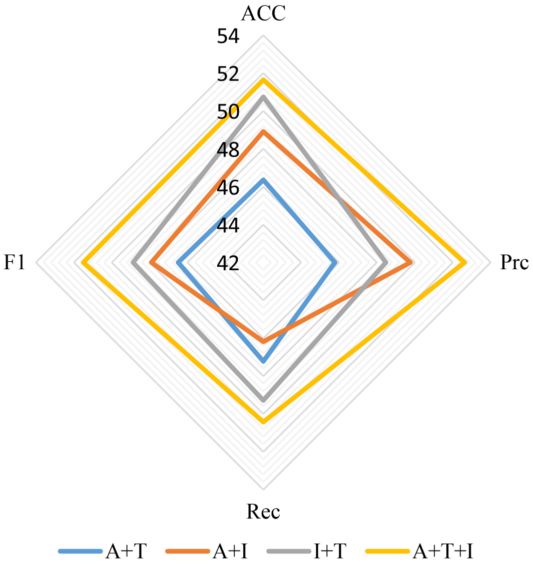

Fig. 7. Experimental results of combination in different modalities based on SIMS. A stands for audio, T stands for text, and I stands for image.

analysis framework, we conducted a series of meticulous ablation experiments on three datasets, namely MOSI, MOSEI, and SIMS. The rationale behind these experiments lies in elucidating the individual impact of these features on the overall model performance across diverse multimodal inputs. This ablation study, designed with meticulous consideration, is poised to offer valuable insights into the nuanced interplay between multimodal multiscale features, thereby enriching our understanding of the proposed sentiment analysis framework's performance across diverse datasets. Specifically, for each dataset, we systematically manipulated the model architecture, progressively excluding the multimodal multiscale features to observe the ensuing impact on sentiment analysis performance. This yielded four experimental configurations as depicted in Fig. 6. As can be seen from Fig. 6, when only a single modality is used, the performance of the model decreases on the three datasets, while using multimodal data will lead to the optimal performance of the model, which illustrates the multimodal multiscale feature learning improves model performance. At the same time, we also noticed that the performance of the speech modality is the lowest, while the performance of the image modality is the highest, which shows that the image modality is more effective for sentiment analysis tasks. In addition, we combined different modalities to show the results of the model on different metrics, as shown in Fig. 7. From Fig. 7, we can see the following results: 1) images play the most important role in sentiment analysis tasks; 2) compared with audio, the text is easier to recognize and analyze by the feature extraction model in sentiment analysis. Our results demonstrate that integrating text, audio, and image data significantly enhances sentiment analysis performance compared to using any single modality. The complementary nature of these modalities allows the model to capture a more comprehensive and nuanced understanding of sentiment expressions.

D. Importance of Fuzzy-DNN Learning

In the proposed MMFDN model, we incorporate fuzzy-deep neural network learning to extract features. Accordingly, in this section, we designed ablation experiments on the Fuzzy-DNN module. Specifically, we replaced the fuzzy layer with an ordinary fully connected layer to obtain model comparison results on three datasets, as shown in Fig. 8. As can be seen in Fig. 8, the four performance metrics consistently indicated a degradation in model performance when fully connected layers replaced the Fuzzy-DNN Learning module. The disparities were evident across all three datasets, underscoring the critical contribution of Fuzzy-DNN to the model's ability to discern and interpret complex, fuzzy sentiments embedded in multimodal data. The observed decline in performance can be attributed to the unique capabilities embedded in the Fuzzy-DNN architecture. Fuzzy logic principles, seamlessly integrated into the neural network, introduce a layer of adaptability and interpretability crucial for handling the inherent ambiguity in sentiment expressions. The ability of Fuzzy-DNN to capture vagueness and uncertainty in emotions appears to be indispensable in the nuanced realm of

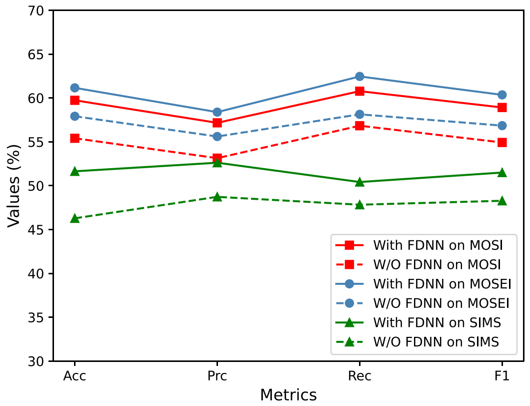

Fig. 8. Experimental results with or without Fuzzy-DNN based on three datasets. "W/O" is the abbreviation of the word "without."

TABLE VII
COMPARISON RESULTS OF MODEL PERFORMANCE WITH AND WITHOUT DAM MODULE

Dataset	Attention	Acc (%)	Prc (%)	Rec (%)	F1 (%)
MOSI	with	**59.74**	**57.17**	**60.78**	**58.92**
MOSI	w/o	51.03	53.53	52.63	53.08
MOSEI	with	**61.16**	**58.39**	**62.46**	**60.36**
MOSEI	w/o	56.65	53.75	57.86	55.73
SIMS	with	**51.64**	**52.62**	**50.42**	**51.50**
SIMS	w/o	46.69	43.84	43.45	43.64

"w/o" is the abbreviation of the word "without."
The bold fonts are the best results for the corresponding metrics to highlight the performance benefits of our model.

sentiment analysis across multimodal datasets. Fully connected layers, while effective in certain tasks, may lack the nuanced adaptability and fuzziness required to decipher the intricacies of human sentiment. The rigid structure of fully connected layers may struggle to capture the subtle transitions and gradations in emotional expressions, leading to a diminished performance compared to the inherently flexible Fuzzy-DNN. The findings underscore the significance of Fuzzy-DNN Learning in our multimodal sentiment analysis framework. The incorporation of fuzzy logic improves the model's ability to handle the inherent uncertainty and ambiguity in sentiment data. Fuzzy logic provides a more flexible approach to dealing with imprecise inputs, which is particularly useful in sentiment analysis where emotions are often not clear-cut.

E. Importance of Dual Attention Mechanism

In this section, our ablation experiments focused on dissecting the influence of the Dual Attention Mechanism (DAM) within our sentiment analysis framework across three distinct datasets. The ablation experiments involved two configurations: One incorporating the DAM and another without, serving as the baseline. These configurations were evaluated on datasets MOSI, MOSEI, and SIMS to ensure a comprehensive assessment of DAM's impact across diverse multimodal sentiments.

The results of ablation experiments on DAM are shown in Table VII. We can see that the performance metrics consistently indicated an enhancement in model performance when DAM is integrated. The improvements were observed across all three datasets, substantiating the effectiveness of the dual attention mechanism in elevating the sentiment analysis capabilities of our model. The observed performance boost with the inclusion of the dual attention mechanism can be attributed to its inherent ability to dynamically highlight and weigh different modalities and regions within the input data. DAM introduces a mechanism for the model to selectively focus on crucial features, effectively improving the model's sensitivity to salient information present in multimodal expressions of sentiment. The attention mechanism is particularly potent in scenarios where certain modalities or regions contribute more significantly to the overall sentiment. DAM facilitates an adaptive and context-aware attention allocation, allowing the model to prioritize relevant information, thus enhancing its understanding of complex sentiment expressions. The positive outcomes of the ablation experiments underscore the importance of incorporating the dual attention mechanism in multimodal sentiment analysis. The implications are substantial, especially in scenarios where certain modalities, such as text, image, or audio, play varying roles in expressing sentiments. DAM equips the model with the capability to discern the varying importance of these modalities dynamically, thereby improving overall sentiment analysis accuracy. To further verify the reliability and accuracy of the model's estimation of the attention mechanism, we selected three test samples from the MOSI dataset, and the output results are shown in Table VIII. It can be seen that after adding the attention mechanism, the model is more accurate and reliable for sentiment analysis. The DAM, which includes both intramodal and intermodal attention, plays a crucial role in enhancing feature representation and improving overall model performance. By focusing on the most relevant features within and across modalities, the model can make more accurate sentiment predictions.

TABLE VIII
EXAMPLES OF COMPARISON RESULTS WITH AND WITHOUT DAM MODULE

Text	Audio	Image	w/o DAM	with DAM	Label
'I thought that film was awful'	blink	gentle	Middle	Negative	Negative
'It looks really good'	blink	smile	Middle	Positive	Positive
'It not bad in the way'	head drop	gentle	Negative	Positive	Positive

"w/o" is the abbreviation of the word "without."

F. Hyperparameter Tuning

We performed a sensitivity analysis to assess how two key parameters, the activation functions and learning rate, affect our model. Table IX compares the model's performance across various activation functions, including ReLU, sigmoid, and tanh, to determine their influence on training dynamics and overall classification outcomes. Our findings suggest that activation functions significantly differ in their effects on convergence rates and performance outcomes. Remarkably, the ReLU activation function stood out, providing the best balance between convergence efficiency and classification precision, surpassing both sigmoid and tanh functions. The effects of different learning

TABLE IX
COMPARISON RESULTS WITH DIFFERENT ACTIVATION FUNCTIONS UNDER THREE DATASETS

Activation function	MOSI	MOSEI	SIMS
sigmoid	51.34	55.35	44.64
tanh	55.36	57.39	46.52
ReLU	**58.92**	**60.36**	**51.50**

The bold fonts are the best results for the corresponding metrics to highlight the performance benefits of our model.

TABLE X
COMPARISON RESULTS WITH DIFFERENT LEARNING RATES UNDER THREE DATASETS

Learning rate	MOSI	MOSEI	SIMS
0.01	45.74	53.28	43.07
0.001	56.19	56.98	45.72
0.0001	**58.92**	**60.36**	**51.50**
0.00001	51.32	52.36	42.53

The bold fonts are the best results for the corresponding metrics to highlight the performance benefits of our model.

rates on the model's performance are detailed in Table X. We found that the learning rate significantly impacts the model's training convergence and ultimate performance. Specifically, lower learning rates led to a more gradual convergence, resulting in steadier training progress and improved generalization capabilities. On the other hand, higher learning rates achieved quicker convergence but increased the risk of training loss fluctuations and overshooting. By analyzing the sensitivity to the learning rate, we pinpointed an optimal range that ensures a good balance between training speed and accuracy in classification.

V. CONCLUSION

In this article, we explore a new multimodal multiscale feature learning model for sentiment analysis based on fuzzy-deep neural network. First, we introduce the specific structure of the proposed model, including the feature representation, the structure of the MMFDN, and the dual attention mechanism. Second, we verify and compare the performance through numerous experiments based on three public datasets. The superior performance of our model on three datasets can be attributed to the effective integration of multimodal data. The fuzzy logic layer plays a crucial role in handling the inherent uncertainty in sentiment expressions, leading to more accurate predictions. Compared to state-of-the-art methods, our model demonstrates significant improvements in accuracy and F1-score. This can be largely credited to the dual attention mechanism, which enhances the model's ability to focus on relevant features across different modalities. Ablation studies reveal that the removal of the fuzzy logic layer results in a noticeable drop in performance, highlighting its importance in managing ambiguous sentiment data. Similarly, excluding the dual attention mechanism leads to less effective feature integration, underscoring its critical role in our approach. In summary, this research endeavors to advance the field of sentiment analysis by embracing the challenges posed by multimodal data and multiscale emotional expressions, offering a novel perspective through the integration of Fuzzy-DNN learning. Through this exploration, we aspire to contribute to the development of more robust sentiment analysis models with broad applications in understanding human emotions across various digital platforms. By establishing a novel sentiment analysis model, our contributions offer an innovative solution for classifying emotions across modalities, promising far-reaching implications for research and applications in this dynamic field. In future work, we aim to advance our research and enhance the capabilities of our sentiment analysis model in several key areas. First, we plan to explore the application of type-2 fuzzy sets to address higher levels of uncertainty and fuzziness. Second, we intend to delve deeper into sophisticated multimodal data fusion techniques. In addition, we are looking to apply our research to real-time sentiment analysis scenarios, such as social media monitoring and customer service automation. This will necessitate optimizing the model for performance and responsiveness to efficiently process streaming data in a real-time setting.

REFERENCES

[1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.

[2] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, 2023.

[3] W. Ding, M. Abdel-Basset, H. Hawash, and W. Pedrycz, "Multimodal infant brain segmentation by fuzzy-informed deep learning," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 4, pp. 1088–1101, Apr. 2022.

[4] C. Kang et al., "A heuristic neural network structure relying on fuzzy logic for images scoring," *IEEE Trans. fuzzy Syst.*, vol. 29, no. 1, pp. 34–45, Jan. 2021.

[5] Ö. Aslan, A. Altan, and R. Hacioğlu, "Level control of blast furnace gas cleaning tank system with fuzzy based gain regulation for model reference adaptive controller," *Processes*, vol. 10, no. 12, 2022, Art. no. 2503.

[6] Ö. Aslan, R. Hacioğlu, and A. Altan, "Pulverized coal injection tank pressure control using fuzzy based gain regulation for model reference adaptive controller," in *Proc. METEC 6th ESTAD*, 2023, pp. 1–8.

[7] M. Pekkaya, Z. Uysal, A. Altan, and S. Karasu, "Artificial intelligence-based evaluation of the factors affecting the sales of an iron and steel company," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 32, no. 1, pp. 51–67, 2024.

[8] Y. Zheng, Z. Xu, and X. Wang, "The fusion of deep learning and fuzzy systems: A state-of-the-art survey," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 8, pp. 2783–2799, Aug. 2022.

[9] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[10] C. Strapparava et al., "Wordnet affect: An affective extension of wordnet," in *Lrec*, vol. 4, no. 40, pp. 1083–1086, 2004.

[11] K. Min, C. Ma, T. Zhao, and H. Li, "BosonNLP: An ensemble approach for word segmentation and POS tagging," in *Proc. Natural Lang. Process. Chin. Computing: 4th CCF Conf.*, 2015, pp. 520–526.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE Proc. IRE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 1–9, 2015.

[15] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 1–11, 2017.

[16] W. Li, F. Qi, M. Tang, and Z. Yu, “Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification,” *Neurocomputing*, vol. 387, pp. 63–77, 2020.

[17] S. Lai, K. Liu, S. He, and J. Zhao, “How to generate a good word embedding,” *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 5–14, Nov./Dec. 2016.

[18] M. Jimenez, J. Aguilar, J. Monsalve-Pulido, and E. Montoya, “An automatic approach of audio feature engineering for the extraction, analysis and selection of descriptors,” *Int. J. Multimedia Inf. Retrieval*, vol. 10, pp. 33–42, 2021.

[19] W. Wang and W. Ding, “Research of improved SVM model based on GA in e-learning emotion classification,” in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Intell. Syst.*, 2012, vol. 2, pp. 919–923.

[20] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, “Online emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues,” *J. Multimodal User Interfaces*, vol. 3, pp. 7–19, 2010.

[21] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP—A collaborative voice analysis repository for speech technologies,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[22] F. Eyben, M. Wöllmer, and B. Schuller, “OpenEAR—Introducing the munich open-source emotion and affect recognition toolkit,” in *Proc. IEEE 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, 2009, pp. 1–6.

[23] M. K. Hasan et al., “UR-FUNNY: A multimodal language dataset for understanding humor,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 2046–2056.

[24] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *The 57th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 527–536.

[25] Z. Luo, H. Xu, and F. Chen, “Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network,” in *AffCon, AAAI*, 2019, pp. 80–87.

[26] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, “Sentiment analysis in non-fixed length audios using a fully convolutional neural network,” *Biomed. Signal Process. Control*, vol. 69, 2021, Art. no. 102946.

[27] X. Zhao et al., “Peak-piloted deep network for facial expression recognition,” in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 425–442.

[28] H. Ding, S. K. Zhou, and R. Chellappa, “FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition,” in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, 2017, pp. 118–126.

[29] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 12, pp. 10790–10797.

[30] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, “Words can shift: Dynamically adjusting word representations using nonverbal behaviors,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 7216–7223.

[31] S. Verma, C. Wang, L. Zhu, and W. Liu, “DeepCU: Integrating both common and unique latent information for multimodal sentiment analysis,” in *Proc. Int. Joint Conf. Artif. Intell. Int. Joint Conferences Artif. Intell. Org.*, 2019, pp. 3627–3634.

[32] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, “Utterance-level multimodal sentiment analysis,” in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.

[33] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “Memory fusion network for multi-view sequential learning,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[34] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, vol. 2019, Art. no. 6558.

[35] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[36] R. Zhang, J. Shen, F. Wei, X. Li, and A. K. Sangaiah, “Medical image classification based on multi-scale non-negative sparse coding,” *Artif. Intell. Med.*, vol. 83, pp. 44–51, 2017.

[37] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, “DS-TransUNet: Dual swin transformer U-net for medical image segmentation,” *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 4005615.

[38] Y. Kong, G. Z. Genchev, X. Wang, H. Zhao, and H. Lu, “Nuclear segmentation in histopathological images using two-stage stacked U-nets with attention mechanism,” *Front. Bioeng. Biotechnol.*, vol. 8, pp. 2296–4185, 2020.

[39] V. Mnih et al., “Recurrent models of visual attention,” *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 1–9, 2014.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[41] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, “MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos,” *CoRR*, 2016, *arXiv:1606.06259v2*.

[42] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[43] W. Yu et al., “CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality,” in *Proc. 58th Annu. meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[44] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” in *Proc. Int. Conf. Representation Learn.*, 2019, pp. 1–20.

[45] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, “Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8992–8999.

[46] D. Hazarika, R. Zimmermann, and S. Poria, “MISA: Modality-invariant and-specific representations for multimodal sentiment analysis,” in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[47] W. Rahman et al., “Integrating multimodal information in large pretrained transformers,” in *Proc. Conf. Assoc. Comput. Linguistics, Meeting*, 2020, vol. 2020, Art. no. 2359.

[48] W. Han, H. Chen, and S. Poria, “Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis,” in *The Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[49] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, “More than a feeling: Accuracy and application of sentiment analysis,” *Int. J. Res. Marketing*, vol. 40, no. 1, pp. 75–87, 2023.

[50] H. Wu, C. Huang, and S. Deng, “Improving aspect-based sentiment analysis with knowledge-aware dependency graph network,” *Inf. Fusion*, vol. 92, pp. 289–299, 2023.