Neurocomputing 501 (2022) 629–639

Contents lists available at ScienceDirect

Neurocomputing

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

Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis

Yuntao Shou [a], Tao Meng [a,*], Wei Ai [a], Sihan Yang [a], Keqin Li [b]

[a] School of Computer and Information Engineering, Central South University of Forestry and Technology, Hunan, China
[b] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Article history:
Received 28 November 2021
Revised 27 May 2022
Accepted 18 June 2022
Available online 23 June 2022
Communicated by Zidong Wang

Keywords:
Dependency parsing
Dialogue emotion recognition
Graph convolution neural network
Self attention mechanism

ABSTRACT

Multimodal Emotion Recognition for Conversation (ERC) is a challenging multi-class classification task that requires recognizing multiple speakers' emotions in text, audio, video, and other modalities. ERC has received considerable attention from researchers due to its potential applications in opinion mining, advertising, and healthcare. However, the syntactic structure characteristics of the text itself have not been considered in this study. Taking into account this, this paper proposes a conversational affective analysis model (DSAGCN) combining dependent syntactic analysis and graph convolutional neural networks. Since words that reflect emotional polarity are usually concentrated exclusively in limited regions, the DSAGCN model first employs a self-attention mechanism to capture the most effective words in the dialogue context and obtain a more accurate vector representation of the emotional semantics. Then, based on speaker relationships and dependent syntactic relationships, the multimodal sentiment relationship graphs are constructed. Finally, a graph convolutional neural network is used to complete the recognition of multimodal emotion. In extensive experiments on two real datasets, IEMOCAP and MELD, the DSAGCN model outperforms the existing models in terms of average accuracy and f1 values for multimodal emotion recognition, especially for emotions such as "happiness" and "anger". Thus, dependent syntactic analysis and self-attention mechanism can enhance the model's ability to understand emotions.

1. Introduction

In recent decades, emotional recognition has had potential applications in developing compassionate robots [1]. With the rapid increase of open dialogue data on social media platforms like Facebook, Twitter, Youtube, and Reddit, more researchers have begun to pay attention to emotional recognition in dialogue [2], and human–computer dialogue systems and diagnose to brain-computer-interfaces have attracted much attention and gradually become a research hotspot in academia and industry. Research on emotion recognition in brain-computer-interfaces and EEG analysis is currently the main way to achieve emotional intelligence [3]. Brain-computer interfaces are mainly used to achieve emotional interaction functions by combining disciplines such as cognitive science and psychology. In a human–computer dialogue system, expressive communication is an important research direction for scholars in relevant fields. Humans can have passionate communication through language and gain emotional comfort [4]. If the human–computer dialogue system wants effective, responsive communication with human beings, it must have enough emotional analysis and judgment ability. Specifically, on the one hand, computers need to identify and judge user emotions, and on the other hand, it needs to incorporate appropriate emotions into the answered messages. Therefore, how to give the machine the ability to understand and express emotion in dialogue is a new challenge in the field of emotional analysis [5].

The key to the breakthrough of ERC includes three significant factors: emotional stimulus (acoustic, visual, audiovisual, text reading, etc...), data collection (EEG recordings, MRI scans, f-NIRS, facial expressions, speech), and the ability of the model to extract data with rich semantic information [6]. We investigate text, video, and audio data collected from individuals in video conversations to identify the speaker's emotion.

In existing ERC task studies, Ghosal et al. [2] processed the sequence contexts composed of dialogue sequentially and modeled the dependencies between speakers by constructing a graph convolutional neural network. Poria et al. [7] used

* Corresponding author.
E-mail addresses: yuntaoshou@csuft.edu.cn (Y. Shou), mengtao@hnu.edu.cn (T. Meng), aiwei@hnu.edu.cn (W. Ai), sihanyang@csuft.edu.cn (S. Yang), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.neucom.2022.06.072
0925-2312/

recurrent neural networks to gather contextual information of the dialogue text. Although the deep learning model is used to extract semantic information in the dialogue context, the syntactic structural features of discourse are ignored. As the basis for understanding the language, the syntactic structure can effectively represent the grammatical structure of the sentences, reveal the relationship between the components of the text, and enhance the ability of the model to understand the emotion of the discourse [8].

Syntactic analysis methods usually contain two classes, syntactic structure analysis, and dependent syntactic analysis. Syntactic structure analysis is mainly used to analyze the phrase structure of sentences, while helpless syntactic analysis focuses on the dependencies between the various components of the sentence. Spatially, dependent syntax trees generated by dependent syntax analysis are a kind of graph data, with words in sentences serving as graph nodes and dependency between words and words as edges. Graph convolutional neural networks are often processed graph structures [9] that capture word-word dependence by GCN by Zhao et al. Lai et al. [10] used the Bi-LSTM and GCN models to extract semantic information on the text and classify emotions in combination with conditional syntactic structure analysis. However, the degree of importance between words is different, reflecting the importance of different emotional feature vectors in the dialogue text. This paper gives different weights between word vectors by introducing attention mechanisms.

In view of this, this paper proposes the multimodal dialogue emotion recognition model combining the dependent syntactic analysis and GCN in accordance with the perspective of the insufficient discourse syntactic structure analysis and weak ability to capture semantic information. First, this paper inputs feature vectors from text, video, and audio modalities into the Bi-LSTM model to obtain feature information of historical contexts, and then adopts a self-attention mechanism to capture the most effective words in dialogue contexts to obtain a more accurate vector representation of emotional semantics. Then multimodal sentiment relation graphs are constructed based on speaker relations and dependent syntactic relations, respectively. Finally, a graph convolutional neural network is used to complete the recognition of multimodal emotions. The DSAGCN proposed in this paper introduces a self-attention mechanism to obtain a more accurate vector representation of emotional semantics. Then, through the dependent syntactic structure, we focus on analyzing the relationship between the components of the sentences, facilitating improving the understanding of the model. The experimental results show that the DSAGCN model is superior to the existing models regarding accuracy and f1 values on two real datasets, IEMOCAP and MELD, especially on emotions such as ‘‘happiness” and ‘‘anger”. Therefore, we believe that DSAGCN can be widely used in emotion analysis, such as human–computer dialogue systems and opinion mining.

The contributions of this paper are as follows:

1. This research proposes a new model called DSAGCN, which combines the adjacency matrix constructed by graph convolutional neural network and dependency syntax analysis, fully considering the dialogue context history of multi-speakers, the dependence relationship between speakers. In addition, the syntactic structure of the dialogue context can be used for emotion recognition in multimodal datasets.
2. Using a multimodal dataset to extract different features of dialogue text, enable the model to learn richer feature information, improve the robustness and introduce a self-attention mechanism to weigh the most emotional feature vectors, reduce the number of network parameters and obtain a more accurate semantic vector representation.
3. In extensive experiments on two real datasets, IEMOCAP and MELD, the DSAGCN model outperforms the existing models in terms of average accuracy and f1 values for multimodal emotion recognition.

The rest of this paper is organized as follows: Section 2 briefly introduces the emotion recognition task; Section 3 defines the problem in mathematical language and describes the data preprocessing; Section 4 describes the research model; Section 5 records the experimental environment and experimental datasets; and finally, Section 2 shows the experimental results and related analysis.

2. Related work

Emotional recognition is an interdisciplinary, multi-domain research task, consisting of psychology, cognitive science, machine learning, and natural language processing [11]. It has many potential applications in a wide range of systems areas, including opinion mining, analyzing the emotional value of customers in business aspects, health care, recommendation systems, education, advertising, an-d more.

In recent years, the problem of emotion recognition in dialogue has attracted much attention in academia and industry due to the rapid increase of open-source dialogue datasets. At present, the study of dialogue emotion recognition is based primarily on the multimodal dataset, which includes three modes: text, audio, and video. In this article, we have likewise adopted these three models.

The emotion recognition problem was initially present-ed with a Convolutional Neural Network (CNN) [12] for text feature extraction, a model independent of context, as it does not use information from contextual utterances. Sukhbaatar et al. [13] proposed that Memnet is an end-to-end memory network where each utterance is input to the network as an input layer, while the memory corresponding to previous utterances is constantly iteratively updated in a multi-hop manner. Finally, the output of the memory network is used for emotion classification. Satt et al. [14] achieved the CNN-LSTM combination model, applied directly to spectral-domain maps, achieving high accuracy.

As the field progresses, the results of scholars in relevant fields on dialogue emotion recognition are mainly based on the constituent sequential discourse of dialogue by recurrent neural networks. Recursive neural networks (RNN) [15] were applied by Poria et al. It relies on spreading contextual and sequential information to the discourse. Tang et al. [16] provided the utterance to the bidirectional gated cycle unit (BiGRU). However, like most current models, they ignored intention modeling, motifs, and personality because of the lack of labels for these aspects of the real datasets. Theoretically, neural networks like long, short-term memory neural networks (LSTM) and gated cycle units (GRU) should propagate long-term contextual information. However, this has not always been the case. Poria et al. [7] used bidirectional LSTM to model the speaker-based context to capture contextual content from the surrounding utterances. However, these contexts cannot perform long-term summary and unweighted effects from context, resulting in huge model bias. This affects the feasibility of the RNN-based model in conversational emotion recognition. Hazarika et al. [17] constructed that CMN, this model utilizes two different GRU to model the discourse context in a historical conversation. Finally, the utterance representation is obtained by providing the current utterance as a query to two separate memory networks. However, this model can only simulate the conversation with the two speakers. ICON [18] is a continuation of CMN, which uses another GRU to connect the output of the individual speaker GRU in CMN to make explicit inter-speaker modeling. The GRU is seen as the memory for

tracking the entire session process. Analogous to CMN, ICON cannot be applied to multimodal datasets. Jiao et al. [19] proposed AGHMN, which first used Bi-GRU to extract features from contextual information, then used Bi-GRU fusion layer to perform feature fusion on historical contextual information, and finally used attention mechanism to update Bi-GRU internal state. AGHMN can balance the importance of historical discourse information and current discourse information. However, AGHMN ignored speaker information. Majumder et al. [20] created that DialogueRNN employs a self-attention mechanism to pool information on all or part of the dialogue with each target utterance. However, the model does not account for the relative positions between the speaker information. We believe that the speaker's information is necessary for contextual information in a distant discourse.

Graph convolutional neural network (GCN) [21,22] is also increasingly popular in solving various graph-based problems, including semi-supervised node classification [21], link prediction [23], recommendation system [24], and others. Ghosal et al. [2] proposed DialogueGCN. To fully model the interactive information between speakers, DialogueGCN detailed the dependencies between speakers using relational GCN. Despite structural improvements in DialogueGCN, examining the relative location of speaker information and other utterances from the target utterances, there are limitations. Entering conversations directly into a two-way recursive module with a pre-trained model offers the opportunity to lose unique features of individual utterances. To alleviate this problem, Choi et al. [25] designed the residual graph Convolutional Neural Network (RGCN), which will generate complex context features for each independent utterance in a ResNet-based internal feature extractor.

Although some progress has been achieved in dialogue emotion recognition research, the above studies have all focused on the semantic and sequential context feature extraction level of dialogue text, ignoring the syntactic structure information of discourse text. However, the syntactic structural information of the text has equally important implications for understanding its emotional polarity with the model. Lai et al. [10] used Bi-LSTM and GCN models, combining semantic information in the text and dependent syntactic trees. However, this way considers the full feature information of the sentence, while the words about emotional polarity are mostly concentrated in confined locations, which will introduce much redundant information that affects the evaluation efficiency of the model.

Therefore, we will introduce a self-attention mechanism, which will effectively alleviate the information redundancy and improve the model's performance.

For the above problems, this paper proposes dialogue textual emotion analysis combining Bi-LSTM, GCN mode-l, syntactic dependency analysis, and self-attention mechanism. The context representation of the dialogue text is obtained through the Bi-LSTM model, and then the self-attention mechanism is introduced to obtain richer emotional features, take the words of the feature as the graph of the initial node, the relationship between words as edges, build the initial state diagram, build the convolutional map based on the syntactic structure, then input the convolutional map together with the initial state map to get the emotional features of the dialogue text in GCN, and finally input the feature into the full connection layer to identify the emotion classification.

3. Preliminary

3.1. Problem definition

The task of this paper is to infer emotional changes in the speaker in a multi-party dialogue system. Suppose that there are M participant $p_1, p_2,..,p_M$ in one conversation, M speakers speak a series of discourse $u_1,..,u_M$ with emotional labels (happy, sad, neutral, angry, excited, and depressed). $u_\lambda = \left(S_\lambda^1, S_\lambda^2, .., S_\lambda^{l_\lambda}\right)$ in chronological order. Among them, S_λ^i is the first i sentence, and l_λ is the total number of words of the speaker $u_\lambda, \lambda \in \{1, .., M\}$. The discourse of the M speakers can be sorted in chronological order as $\left(\delta_1, .., \delta_{l_1+..+\delta_{l_M}}\right)$, where $\delta_j \in \{u_1, u_2, .., u_M\}$.

The model constructed in this paper takes emotion labels requiring classified utterances δ_i as input to the network, to obtain historical information about the dialogue context, set a context window of size K to record everyone's historical utterances. Among them, the $H_\lambda \in \{p_1, p_2, .., p_M\}$. The definition of H_λ is as follows:

$$H_\lambda = \{\delta_i | i \in [t-K, t-1], \delta_i \in u_\lambda, |H_\lambda| \leqslant K\} \quad (1)$$

3.2. Multimodal feature extraction

The first step of our algorithm is tantamount to extracting the multimodal features of all utterances in the conversational dataset. For each discourse, we extract the features of these discourses from three different modes of text, audio, and video. The feature extraction procedure for each mode is characterized below.

3.2.1. Textual features extraction

Text features of the dialogue context are significant grounds for identifying the emotional changes of the speak-er. Convolutional neural network (CNN) proposed by Kim et al. [12] was utilized to perform feature extraction of conversational discourse text. Convolutional neural networks can effectively learn highly abstract semantic features from the words that constitute sentences. This paper requires a convolutional layer, a maximum pooling layer, and a fully connected layer to obtain a characteristic representation of the dialogue discourse text. The input to the network is comprised of 300-dimensional pre-trained word embeddings. The convolutional layer is separately composed of three convolutional filters of sizes 3,4,5. Each with 50 feature maps and the complex features of the dialogue discourse text will be maximized together. We use these three filters to perform one-dimensional convolutional kernel operations, and then output them to the maximum pooling layer at its maximum pooling operations. It is then fed into a 100-dimensional fully-connected layer by a nonlinear transformation using the activation Relu function, and the final output forms a characteristic representation of text utterance. This network is trained with effective labels.

3.2.2. Audio feature extraction

Audio plays a significant role in identifying the speak-er's emotions. To extract audio context features, this paper first stores the speaker's audio in the video as a 16-bit PCM WAV file and then uses the openSMILE [26] open-source toolkit for audio contextual feature extraction. An openSMILE enables high-dimensional audio vectors composed of MFCC, Mel-spectra, pitch, loudness, et al. In particular, we will use the IS13_ComParE1 extractor[1] from openSMILE to obtain a 6373-dimensional feature vector for each sentence. Furthermore, to constrain the feature vectors in the range [0, 1], we use Min–Max normalization to scale the feature vectors. Since the dimension of the obtained audio vectors is too high, it is easy to cause the model to overfit, so we use the fully connected layer to map the audio vectors to the 100-dimensional feature vector α_u. The low-dimensional feature vector $\alpha_u \in \mathbb{R}^{d_\alpha}$ is the final audio vector.

[1] http://audeering.com/technology/opensmile

3.2.3. Visual feature extraction

Visual features, such as movements, also reflect changes in a person's emotional characteristics. Therefore, we capture visual features from the conversational video using the 3D-CNN algorithm [27]. 3D-CNN serves to understand human facial expressions, such as frowning or smiling. It needs to be able to extract relevant features from each frame of the image. The input to the network is a vector of size (c,h,w,f). Among these, c represents the number of channels of the image, the video used in this paper has three channels of RGB. h and w represent the height and width of the image per frame, and f represents the total number of frames in the video. The 3D-CNN network consists of three convolutional blocks. Each convolutional block contains two convolutional layers of size 5 $\times$ 5 and a maximum pooling layer of size 3 $\times$ 3. For the convolution operation, a 3D filter of dimension (f_o,f_i,f_h,f_w,f_d) is used. Among them, f_o,f_i,f_h,f_w,f_d represent the feature map, the number of input channels, the image width, the image height and the depth of the 3D filter, respectively. A max pooling operation is used after the output of the convolutional layer. After convolution and max pooling operations, it is input into the ReLU activation function for nonlinear transformation. Finally, the resulting final feature is mapped to a dense fully connected layer dimension d_v, whose activation acts as a visual feature $v_u \in \mathbb{R}^{d_v}$. Among them, $d_v = 512$.

4. Methodology

4.1. Our model

In the multimodal dialogue emotion recognition problem, this paper presents the conversational-dependent syntactic analysis and the graph convolutional Neural Network (DSA-GCN) framework, with models specifically consisting of Bi-LSTM, self-attention mechanisms, dependent syntactic structure, and GCN. First, the dialogue text is coded into a word vector through the Word2Vec model, input the video features, audio features, and the word vectors corresponding to the dialogue text into Bi-LSTM to get the historical context information features of the speaker. In order to reflect the importance of different emotional feature vectors in the conversation text, a self-attention mechanism is introduced to capture the most effective words in the dialogue context, so as to obtain a more accurate vector representation of emotional semantics. Then, based on speaker relations and dependent syntactic relations, respectively, the multimodal sentiment relation graphs are constructed and input to the GCN layer to aggregate emotional semantic information. Finally, video features, audio features, and word vectors that pass through the GCN layers are input to the emotion classifier consisting of softmax layers to obtain discrete sentiment categories. The overall architecture diagram of the DSAGCN model proposed in this paper is shown in Fig. 1.

4.1.1. Sequential contextual features extracted

The quality of the text feature representation has important effects on downstream tasks, and the Word2Vec model [28] is in a position to learn rich semantic knowledge in an unsupervised manner from a large number of corpora. This paper uses the Word2Vec model to generate word vectors with rich languages. Text ℓ_i is prearranged using a pretrained word vector model to get. $\ell_i = \left\{\ell_i^1, \ell_i^2, \ldots, \ell_i^j, \ldots, \ell_i^n\right\}$, Among them, $\ell_i^j \in \mathbb{R}^{d_w}$, The n indicates the sentence length, and the d_w word vector indicates dimension. We set $d_w = 100$. Since the conversation is held in order, the context information passes it in that order. Bi-LSTM is comprised of forwarding and reverse LSTM that can better learn the contextual information of the sentence. We, therefore, input ℓ_i into Bi-LSTM for feature extraction, and the resulting hidden layer feature h_i is spliced from features extracted from forward and reverse LSTM. The specific formula is as follows:

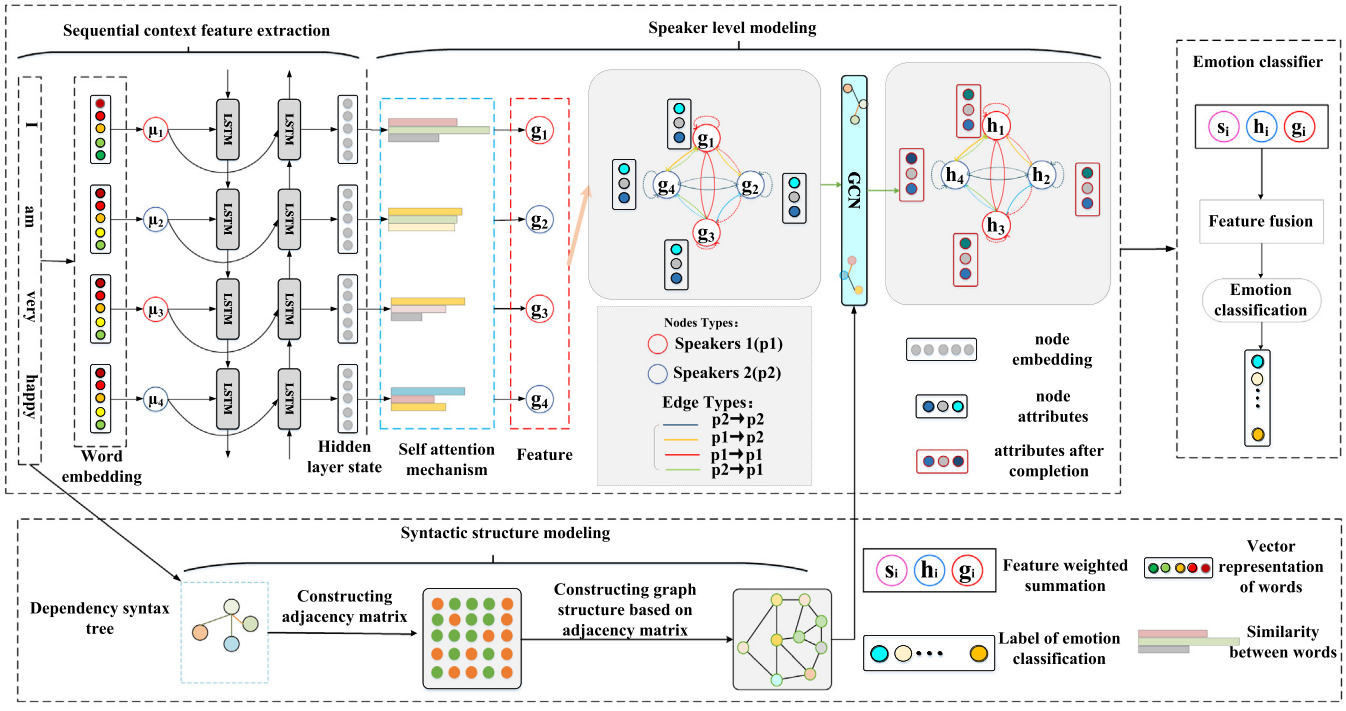


Fig. 1. DSAGCN architecture consists of sequential contextal features extraction, self-attention mechanisms, speaker relationship modeling, dependent syntactic analysis, and emotion classification.

$$\begin{bmatrix} \widetilde{C}_t \\ O_t \\ j_t \\ f_t \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} W_T \begin{bmatrix} \ell_i^t \\ h_i^{t-1} \end{bmatrix} \tag{2}$$

$$C_t = C_t \odot j_t + C_{t-1} \odot f_t \tag{3}$$

$$h_i^t = O_t \odot \tanh(C_t) \tag{4}$$

where, $h_i = \left\{ h_i^j | h_i^j \in \mathbb{R}^{d_l}, j = 1, 2, \ldots, n \right\}, d_l$ represents the number of hidden layer cells. We set the number of hidden layer units of LSTM to 400 in our experiments. Since Bi-LSTM is composed of forward and reverse LSTMs, the final h_i is an 800-dimensional feature vector. l_i^t represents input at t moment, W_T the weight matrix, σ the sigmoid activation function and f_t, j_t, O_t represents the forgetting, input and output gates at t moment, respectively, $\odot$ represents the point multiplication operation, and the tanh function indicates the hyperbolic tangent activation function.

4.1.2. Self-attention mechanism

To reflect the differences in the importance of the emotional different feature vectors in the dialogue text, enhance the emotional semantic representation of the context, and reduce the computation of the network, this study introduces a self-attention mechanism [29]. In conversational texts, words that embody emotional characteristics are usually limited to the local location of the sentence, and using self-attention mechanisms can capture the most emotional words in the text and give it a high attention weight. Although Bi-LSTM can catch certain long-distance text contextual features, it also causes information loss as recursive networks deepen, and its ability to capture long-distance text contextual features weakens. The attention mechanism is essentially the mapping of the input Query (Q) to a series of key-value pairs (Key (K)). Values (V)), and the addition of the attention mechanism reduces the loss of information in the recursive process.

In order to realize the important difference of different emotional feature vectors in the dialogue text, this study designed the following attention mechanism to obtain the semantic vector representation S_i:

$$S_i = \sum_{t=1}^{T} a_{t,i} S_i^t \tag{5}$$

$$a_{t,i} = \text{softmax}\left(\frac{\exp(e_{t,i})}{\sum_{k=1}^{k} \exp(e_{t,k})} \right) \tag{6}$$

$$e_{t,i} = \frac{Z_k S_i^t}{\sqrt{|Z_k|}} \tag{7}$$

where, t represents moment t in T moments, and i represents the semantic vector sequence number of the text sequence in the network. Therefore, $a_{t,i}$ is the importance weight to the semantic vector S_i^t at t, S_i^t is the semantic vector at t, Z_k is the attention vector of the network pattern at t, and $e_{t,i}$ is the similarity between the semantic vector S_i^t and the attention vector of the network pattern at t.The architecture of the self-attention mechanism proposed in this paper is shown in Fig. 2. The three modal features of text, audio and video are first input to the Bi-LSTM for semantic information extraction, then the obtained feature vectors are passed through the softmax function to obtain the attention scores, and finally the attention scores are weighted and summed with the obtained feature vectors to obtain the final semantic vector representation S_i.

4.1.3. Speaker relationship modeling

We model the dependencies between speakers and the-mselves to capture speaker-level dependent context information in a conversation. We construct a directed graph to characterize the emotional interaction between the speaker and itself, and input it into the graph convolutional neural network to obtain a feature representation containing speaker-level context information.

This study was used to represent relationships between interlocutors by constructing a graph $G = \{V, E, R, W\}$. V represents the set of nodes, and E represents the set of edges. Each discourse is represented as a node $V_i \in V, i = 1, 2, \ldots N$, Each vertex V_i is represented by the sequential context-encoded feature vector g_i, and the edge $V_{ij} \in E, r \in R$ between node V_i and node V_j represents the relationship type between nodes.

For edge weights, assuming Y edges, and a self-attenti-on mechanism based on text similarity is used to set the edge weights so that the sum of the input edges is one, considering the past m sentences $V_{t-1}, \ldots, V_{t-m}$ of node t and the later n sentence $V_{t+1}, \ldots, V_{t+n}$. The weight calculation formula is as follows:

$$\begin{aligned} &w_{ts} = \text{softmax}\left(g_t^T W \left[g_{t-m}, \ldots, g_{t+n}\right]\right), \\ &s = t-m, \ldots, t+n \end{aligned} \tag{8}$$

Graph convolutional neural network aggregates the local neighbor feature information about each node, and converts the context feature vector irrelevant to the speaker g_i into a vector representation related to the speaker u_i through a two-step convolution operation. The calculation formula is as follows:

$$\begin{aligned} &u_i^{(1)} = \sigma\left(\sum_{r \in R_s \in Y_i^r} \sum_{c_{t,r}}^{w_{ts}} W_r^{(1)} g_s + w_{ut} W_0^{(1)} g_t \right), \\ &i = 1, 2, \ldots, Y \end{aligned} \tag{9}$$

$$\begin{aligned} &u_i^{(2)} = \sigma\left(\sum_{s \in Y_i} W^{(2)} u_i^{(1)} + W_0^{(2)} u_i^{(1)} \right), \\ &t = 1, 2, \ldots, Y \end{aligned} \tag{10}$$

where, σ is set to the ReLU activation function $W_0^{(1)}, W_r^{(2)}, W_0^{(2)}, W^2$ as the transformation parameter, and $W_{ut}, W_{ts} \in W, Y_i^r$ represents the adjacency index of the node V_t in the relational $r \in R, C_{t,r} \in Y_i^r$. Eqs. (9) and (10) effectively aggregate the local neighborhood speaker information of each sentence node.

4.1.4. Modeling by dependent syntactic structure

We not only consider the dependencies between sequential context information and discourse level, but also combine the dependent syntactic structure of sentences and introduce GCN to study the conversational emotion recognition problem. A dialogue emotional text graph based on the dependent syntax tree was constructed using $G = \{V, E\}$. V represents the set of nodes, the set of words. E represents the set of edges, the set of dependency between words and words. For example, "I am very happy," the adjacency matrix constructed based on a dependent syntactic tree is shown in Fig. 3. Each word in the sentence is adjacent to itself, namely, the diagonal elements in the adjacency matrix are all 1, based on the dependency in the dependent syntax tree, if a word in the sentence and other words have the dependency, the corresponding position in the adjacency matrix is 1, otherwise 0. Therefore, we obtain a node position in a sparse adjacency matrix A, graph in one-to-one correspondence to the positions of elements in S_i obtained based on the self-attention mechanism. Then, through the graph based on the adjacency matrix A, the graph con-

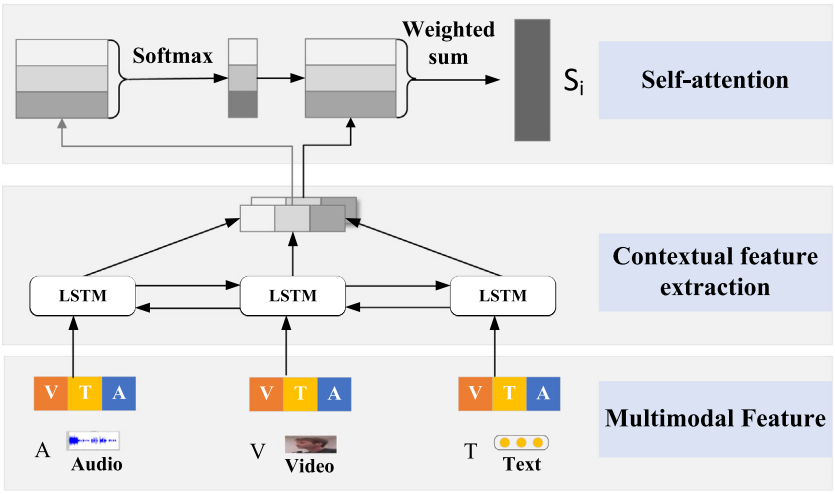
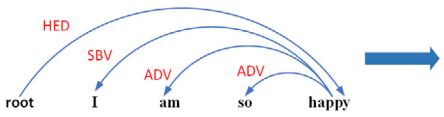


Fig. 2. Multimodal self-attention mechanism architecture.



	I	am	so	happy
I	1	0	0	1
am	0	1	0	1
so	0	0	1	1
happy	1	1	1	1

Fig. 3. Depdependent syntactic structure analysis and its corresponding adjacency matrix. HED stands for a core relationship, SBV stands for a subject-predicate relationship, and ADV stands for an adverbial relationship.

volutional neural network operates the convolution features, and finally obtains the feature V_i. The specific formula is as follows:

$$V_i = \mathrm{ReLU}\left(\widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}}S_iW_C\right) \tag{11}$$

among these, $\widetilde{D}$ is the moment of the adjacency matrix A, $\widetilde{D} = \sum_j A_{ij}$, ReLU function represents the activation function, and W_C is the weights of neurons in the layers in the graph convolutional neural network.

4.1.5. Emotion classification

The emotion classifier first connects the feature vector g_i containing sequence context information, the feature vector u_i from the speaker relationship modeling and the feature vector V_i based on the dependent syntactic structure modeling, then obtains the new dialogue text feature representation through the similarity-based attention mechanism, and finally classifies the discourse using the full connection layer to obtain the corresponding emotion category label P_i. In practical applications, different real datasets usually have different sentiment class labels. For example, in the IEMOCAP real dataset, the sentiment class label $P_i = 7$ obtained after going through the sentiment classifier is 6, and in the MELD real dataset, the sentiment class label $P_i = 7$ obtained after going through the emotion classifier.

As shown in Eq. (12), the extracted context feature vector g_i is connected to the feature vector u_i from the speaker relationship modeling and the feature vector V_i from the dependent syntactic structure modeling, and the vector of h_i is expressed as follows:

$$h_i = [g_i, u_i, V_i] \tag{12}$$

As shown in Eqs. (13) and (14), obtaining the final discourse representation $\widetilde{h}_i$ from the connection using the similarity-based attention mechanism. The $\widetilde{h}_i$ is defined as follows:

$$\beta_i = \mathrm{softmax}\left(h_i^T W_\beta [h_1, h_2, \ldots, h_N]\right) \tag{13}$$

$$\tilde{h}_i = \beta_i [h_1, h_2, \ldots, h_N]^T \tag{14}$$

Finally, the resulting sentence feature representation $\widetilde{h}_i$ after graph convolution operation was input into a layer of the fully-connected neural network, after a nonlinear transformation of the Relu function, and then input into the softmax layer to obtain the emotional tag $\hat{y}_i$ with maximum probability, where y_i is a number in the range [0,1]. The representation of $\hat{y}_i$ is as follows:

$$l_i = \mathrm{ReLU}\left(W_l\tilde{h}_i + b_l\right) \tag{15}$$

$$P_i = softmax(Wl_i + b) \tag{16}$$

$$\hat{y}_i = \underset{t}{\mathrm{argmax}}\,(P_i[t]) \tag{17}$$

4.1.6. Model training

The DSAGCN model uses a cross-entropy loss function to optimize the learning effect for each emotion. The function is defined as follows:

$$\mathscr{L}^i_m(\theta) = \left(y^i_m \log\left(\hat{y}^i_m\right) + \left(1 - y^i_m\right)\left(1 - \log\left(\hat{y}^i_m\right)\right)\right) \tag{18}$$

where θ is a learnable parameter in the network, $\hat{y}^i_m$ denotes the prediction probability of the i-th emotional label, and y^i_m denotes the true class of the i-th emotional label. Generally, the smaller $L^i_m(\theta)$, the closer the probability distribution between the predicted emotional value of the model and the real emotion, the better the effect of emotion recognition.

Based on the above-mentioned loss function, the learning goal of the DSAGCN model is to obtain the minimum loss for various emotion recognition in the training sample data. The learning objective function is defined as follows:

$$\min_\theta \sum_{m=1}^{N_1}\sum_{i=1}^{N_2} \mathscr{L}^i_m(\theta) \tag{19}$$

where, N_1 and N_2 represent the number of training samples and emotion label categories, respectively. The optimal network parameters are obtained when the DSAGCN model obtains the minimum loss value for various emotion recognition training data.

5. Experimental setting

The research environment is based on the Ubuntu 18.04 operating system, using the programming language Python 3.8, deep learning framework Pytorch 1.9.1, hardware for two servers with Tesla P4, the memory of 16G for algorithm comparison experiments.

5.1. Datasets used

We used two conversational emotion datasets, IEMOCAP [30] and MELD [31], to evaluate the effect of the DSAGCN algorithm. We divided these two datasets into training and test sets, with an approximate ratio of 80:20. In the IEMOCAP dataset, the training and validation sets contain 120 dialogues and 5,810 dialogue texts, and the testing sets have 31 dialogues and 1,623 dialogue texts. In the MELD dataset, the training and validation sets contain 1,153 dialogues and 11,098 dialogues, and the testing sets contain 210 dialogues and 2,610 dialogues. Table 1 displays the distribution of samples for both the training and testing sets.

1. The IEMOCCAP datasets contain ten speakers communicating in a two-way conversation, providing thr-ee modalities: text, audio, and video, with 7,433 text utterances and approximately 12 h of audio and video. Each video includes two dialogues with each other which are divided into discourse. Each discourse has six different emotional labels, namely, happiness, sadness, neutrality, anger, excitement, an-d frustration. The final emotion category for each utterance was determined jointly by six evaluators.
2. The MELD dataset, from the Friends series, unlike the IEMOCAP dataset, has multiple speakers involved in conversations and is owned by the multi-conversation dataset. Its training, validation, and test sets have 1,039, 114, and 280 conversations, respectively. Each discourse comes with one of the seven emotional labels, namely anger, disgust, sadness, joy, neutrality, surprise, and fear. The final emotion category for each utterance was determined jointly by five evaluators.

5.2. Comparison algorithm

We compare our model with the following baselines:

CNN: CNN proposed by Kim et al. [12] represents the criteria of text classification that does not consider text context, the context of the speech, or multimodal data.

bc-LSTM: The characteristic representation of the bidirectional LSTM capturing the discourse context from the surrounding discourse, proposed by Poria et al. [7], does not consider the inter-speaker relationship as it does not model the inter-speaker dependence.

CMN: CMN proposed by Hazarika et al. [17] used for GRU to model the historical context of dialogue discourse and input the current discourse into the memory network to obtain a characteristic representation of the discourse context. However, the modification model can only be used to model binary dialogue relationships.

DialogueRNN: DialogueRNN, proposed by Majumd-er et al. [20], is a recurrent network that uses GRU to track individual speaker states during a conversation while introducing self-attention mechanisms that capture attention scores from the textual context, taking into account the relevance of emotional context discourse, and performing emotion classification based on this information. The model can be applied to a multimodal dataset.

DialogueGCN: DialogueGCN proposed by Ghosal et al. [2] details the relationship between speaker dependence and self-dependence that can be applied to multi-party datasets. Combine sentence-as-sentence encoding with speaker coding to enhance the representation of the dialogue context.

AGHMN: The AGHMN proposed by Jiao et al. [19] adopted an attention-gated hierarchical memory network. AGHMN first used Bi-GRU to extract bidirectional contextual history information and then used the attention mechanism to calculate the attention score of the contextual information. Finally, the internal state of BiGRU was updated with the obtained attention score. AGHMN can balance contextual semantic information from recent memory and distant memory.

Table 1
Division of training set and test set in two conversational emotion datasets.

Dataset	Partition	Utterance Count	Dialogue Count
IEMOCAP	train + val	5810	120
	test	1623	31
MELD	train_val	11098	1153
	test	2610	280

5.3. Evaluation metrics

To verify the validity of the DSAGCN model on the IEMOCAP and MELD datasets, we use accuracy and F-score metrics for evaluation, respectively.

The accuracy was defined as follows:

$$\text{Accuracy} = \frac{\sum_{j}^{\vartheta_1} A_j^T}{\sum_{i=1}^{\vartheta_2} x_i} \tag{20}$$

where ϑ_1, ϑ_2 are the number of samples in the testing datasets and samples whose emotion is correctly predicted by the model, respectively. x_i is the i-th sample in the testing datasets, and A_j^T represents the correct sentiment prediction value for the j-th sample. Generally, a larger accuracy indicates a better prediction of the model.

The F-score integrates precision and recall. Therefore, in this paper, the F-score is also chosen as another metric to evaluate the model's validity. The F-score is defined as follows:

$$F-\text{score} = 2 \cdot \frac{\text{precision}(E_p, E_t) \times \text{recall}(E_p, E_t)}{\text{precision}(E_p, E_t) + \text{recall}(E_p, E_t)} \tag{21}$$

and

$$\text{precision}(E_p, E_t) = \frac{|E_p \bigcap E_t|}{|E_P|} \tag{22}$$

$$\text{recall}(E_p, E_t) = \frac{|E_p \bigcap E_t|}{|E_P|} \tag{23}$$

where E_P represents the predicted emotion label category, E_t represents the real emotion label category, precision (E_p, E_t) denotes precision, recall(E_p, E_t) denotes recall. Generally, the higher the F-score of the model, the better the effect of emotion recognition.

6. Results

To avoid experimental coincidence, we randomly performed 10 training and testing sessions according to the data set division ratio in Table 1, and then took the average value as the final accuracy and f1 value for each emotion. Finally, the accuracy and f1 value

for each emotion are averaged to obtain the average accuracy and f1 value for each algorithm.

By comparing the classification accuracy and f1 values of each discrete emotion label in Tables 2 and 3, we find that the classification accuracy of DSAGCN for "happy" and "anger" emotion labels are 60.1% and 52.2%, respectively, and the f1 values are 62.6% and 46.9% respectively. And the accuracy and f1 values are both about 20% to 30% higher than other models, significantly outperforming existing models. We believe this is because the self-attention mechanism serves is used to capture the semantic information, while the dependent syntactic analysis enhances the model's ability to understand emotions. Meanwhile, in the classification accuracy and f1 values of "happy" discrete emotion, the effect of the DialogueGCN model and AGHMN model is second, the classification accuracy is 45.7% and 42.7%, and the f1 value is 47.7% and 51.1%, respectively, which is about 15% lower than DSAGCN. The classification accuracy and f1 values of the other models are close, between 27% and 35%. In the classification of "happy" discrete emotions, their performance is much worse than DSAGCN, DialogueGCN, and AGHMN. In terms of classification accuracy and f1 value for the "anger" discrete emotion, the DialogueRNN model is slightly less effective than DSAGCN, with the classification accuracy and f1 value of 41.0% and 41.5% respectively. The classification accuracy and f1 values of the other models are close to but lower than the DialogueRNN model. On the "fear" and "disgust" sentiment labels, AGHMN achieves the best results among all models with classification accuracies of 9.8% and 14.0%, and f1 values of 10.6% and 16.4%, respectively. The classification accuracy and f1 value of DSAGCN and other models do not exceed 10%. All models, including AGHMN, performed poorly on the "fear" and "disgust" sentiment labels, and the classification results were unreliable. However, on the "fear" and "disgust" emotion labels, the classification accuracy and f1 values of DSAGCN and existing models are not more than 10%, showing poor performance. We believe this is because the MELD dataset used in this paper comprises video clips from the Web. In the video clips, the speakers rarely show "fear" or "disgust" emotions. As shown in Table 4, the number of "fear" and "disgust" sentiment labels are 50 and 68, respectively, which are far less than the number of other sentiment labels, which will lead to severe data imbalance problem. In the vast majority of cases, the model is usually biased towards learning an unbiased representation of the majority class samples and treats the minority class samples as outliers in the data. It will result in the model not being adequately trained on the minority class sentiment, showing an underfitting state. Furthermore, the speaker is always subtle in expressing such emotions, and the model can easily identify them as "neutral" or "sadness." Since the model proposed in this paper and other models did not consider the problem of imbalanced data distribution, the model's classification accuracy on the "fear" and "disgust" sentiment labels is relatively poor.

As shown in Table 2, the DSAGCN model has the best results on the IEMOCAP dataset, with an average accuracy of 63.5% and an average f1 value of 61.7%, which is 4% to 13% higher than the other models. The effect of the AGHMN model is second, with the average accuracy and f1 value of 61.9% and 61.8%, respectively. DialogueGCN, bc-LSTM, CMN, and DialogueRNN models have similar effects but are slightly worse than the AGHMN model. CNN model has the worst effect, and the average accuracy and f1 value are only 48.0% and 48.1%, respectively. As shown in Table 3, the DSAGCN model has the best results on the MELD dataset, with an average accuracy of 60.9% and an average f1 value of 58.7%, which is 2% to 6% higher than the other models. The effect of AGHMN is second, the average accuracy and f1 values are 58.8% and 57.0%, respectively, which is slightly worse than the DSAGCN model. The other

Table 2
Compare various baseline methods on the IEMOCAP dataset; bold font indicates the best performance; Acc = Accuracy; Average(w) = Weighted average.

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CNN	27.7	29.8	57.1	53.9	34.3	40.1	61.1	52.4	46.1	50.0	62.9	55.7	48.0	48.1
bc-LSTM	29.1	34.4	57.1	60.8	54.1	51.8	57.1	56.7	51.1	57.9	67.1	58.9	55.2	54.9
CMN	25.0	30.3	55.9	62.4	52.8	52.3	61.7	59.8	55.5	60.2	**71.1**	60.6	56.5	56.1
DialogueRNN	33.5	35.4	69.0	68.8	54.1	54.7	**67.1**	61.1	55.9	60.4	62.9	60.3	58.3	58.1
DialogueGCN	45.7	47.7	**86.9**	**84.4**	41.9	48.5	61.5	**62.2**	**72.4**	69.3	51.5	56.6	59.0	56.1
AGHMN	42.7	51.1	63.4	68.0	**61.3**	**57.4**	61.9	61.8	67.5	70.5	64.1	60.5	61.9	61.8
DSAGCN	**60.1**	**62.6**	84.8	82.3	44.5	47.5	63.7	59.6	69.3	**71.5**	54.8	**62.1**	**63.5**	**61.7**

Table 3
Compare various baseline methods on the MELD dataset; bold font indicates the best performance; Acc = Accuracy; Average(w) = Weighted average.

Methods	MELD															
	Neutral		Surprise		Fear		Sadness		Joy		Disgust		Anger		**Average(w)**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CNN	76.2	74.9	43.3	45.5	4.6	3.7	18.2	21.1	46.1	49.4	8.9	8.3	35.3	34.5	54.3	55.0
bc-LSTM	78.4	73.8	46.8	47.7	3.8	5.4	22.4	25.1	51.6	**51.3**	4.3	5.2	36.7	38.4	57.5	55.9
DialogueRNN	72.1	73.5	**54.4**	**49.4**	1.6	1.2	23.9	23.8	52.0	50.7	1.5	1.7	41.0	41.5	56.1	55.9
DialogueGCN	70.3	72.1	42.4	41.7	3.0	2.8	20.9	21.8	44.7	44.2	6.5	6.7	39.0	36.5	54.9	54.7
AGHMN	**80.3**	**75.1**	53.7	49.1	**9.8**	**10.6**	19.7	**25.5**	50.5	51.1	**14.0**	**16.4**	33.9	38.2	58.8	57.0
DSAGCN	76.7	74.4	48.6	45.5	5.2	4.8	**24.4**	22.1	**52.5**	49.6	7.4	8.7	**52.2**	**46.9**	**60.9**	**58.7**

Table 4
The distribution of each label category in the MELD dataset.

Dataset	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger
MELD	1256	281	50	208	402	68	345

models' average accuracy and f1 values are closer to each other and lower than the DSAGCN and AGHMN models.

There are significant differences in performance between DSAGCN and other comparison algorithms, and we believe that the most important reason is the different nature of the model. Both DSAGCN and DialogueGCN attempted to model the dependencies between speakers, while other comparison algorithms considered only feature-encoding of the sequential context. This is a major problem with the comparison algorithms because what the spe-aker says has an essential impact on the listener. In addition, DSAGCN and DialogueGCN used graph convolution operations to simulate the interaction process between speakers, which further obtained the position-encoding information of the utterance context. At the same time, AGHMN utilized the attention mechanism to update the internal state of Bi-GRU to obtain the position-encoding information of the utterance context. However, other comparison models ignored the above problems, which is an important reason for the poor performance of the comparison models because the semantic information of the same word in different positions may vary significantly.

As for the performance difference between DSAGCN and DialogueGCN, it is considered that part of the dependent syntactic structure modeling causes it.DialogueGCN does not consider the syntactic structural features of the sentence. However, the syntactic structure can reveal dependencies between each sentence component, benefiting the model to learn richer feature information. And, words that highlight emotion often appear only locally in the dialogue context, while Dialogue considers the detailed location of the dialogue context, adding too much redundant information. In contrast, DSAGCN models the sentence structure by introducing a syntactic dependent analysis, which helps to enhance the model's ability to understand emotions. The self attention mechanism is introduced to capture the most expressive words in the dialogue context and obtain richer semantic information.

As for the difference in performance between DSAGCN and AGHMN, we believe that AGHMN ignored speaker information and relied on syntactic structure modeling. Identifying which speaker the context belongs to can provide the model with semantic information about the speak-er's emotional changes during the interaction process to better classify the emotional labels.

The above experimental results show that the DSAGCN model proposed in this paper outperforms the existing models on both IEMOCAP and MELD datasets, syntactic dependency relationship, attention mechanism, and the use of graph convolutional neural network to model the dependencies between speakers can effectively improve the emotion recognition ability of the model.

7. Discussion and conclusion

This paper presents the analysis of the predicted labels and finds the model indistinguishable for similar emotion category labels. In the confusion matrix, we found that our model misclassified some with "frustrated" or "happy" labels into "neutral" labels, possibly due to the small differences between these emotional labels. We believe that this ambiguity can be disambiguated by increasing the dataset.

The innovation of our model is to introduce dependent syntactic analysis and self-attention mechanisms, and investigate the effect of these two parts on the model effect. We evaluate their impact on the model performance by removing both parts.

As shown in Table 5, dependent syntactic structure modeling has a significant improvement in the model understanding emotional ability, and without this part, the model performance will decrease by 3.7%. We believe that a full understanding of the dependencies between the various components contributes to better semantic information.

Table 5
Results of the ablation studies on the IEMOCAP dataset.

Dependency parsing	Self-attention mechanism	F1
–	+	58.0
+	–	59.3
+	+	61.7

Self-attention mechanisms also have effects on model performance, but less than context-dependent syntactic analysis. Its absence causes a 2.4% drop in performance. We argue that because the self-attention mechanism can capture the most expressive words in a sentence.

In conclusion, in this paper, a network structure based on a graph convolutional neural network and dependent syntactic analysis for categorical recognition of emotions. Compared with other comparison algorithms, the method presented here considers each input discourse for the degree of similarity between word-to-word, which provides better semantic vector representation and introduces dependent syntactic structure analysis that enhances the mo-del's ability to comprehend discourse emotion. On the IEMOCAP and MELD data sets, the average accuracy of DSAGCN is 63.5% and 60.9%, respectively, and the f1 values are 61.7% and 58.7%, respectively. The average accuracy and f1 values of DSAGCN outperform existing models, especially for emotions such as "happiness" and "anger". Therefore, we believe that DSAGCN can be widely used in emotion analysis, such as human–computer dialogue systems and opinion mining. In future research work, we consider extracting to obtain a better utterance feature representation by transformer, and we also think of introducing self-attention mechanisms within the GCN to improve model performance. In addition, due to the serious problem of class imbalance in the MELD dataset, we will also make full use of the data imbalance solution in future research work to improve the classification accuracy of the model on the minority class.

CRediT authorship contribution statement

Yuntao Shou: Conceptualization, Methodology. **Tao Meng:** Data curation, Writing - original draft. **Wei Ai:** Validation, Supervision. **Sihan Yang:** Software. **Keqin Li:** 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.

Acknowledgments

We thank the reviewers so much for their professional comments and advice. This work is supported by National Natural Science Foundation of China (Grant No. 61802444); the Research Foundation of Education Bureau of Hunan Province of China (Grant No. 20B625, No. 18B196); the Research on Local Community Structure Detection Algorithms in Complex Networks (Grant No. 2020YJ009).

References

[1] Z. Lian, B. Liu, J. Tao, Ctnet: Conversational transformer network for emotion recognition, IEEE/ACM Trans. Audio, Speech, Language Process. 29 (2021) 985–1000.
[2] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, in:

EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2020..
[3] S. Aydın, S. Demirtaş, S. Yetkin, Cortical correlations in wavelet domain for estimation of emotional dysfunctions, Neural Comput. Appl. 30 (4) (2018) 1085–1094.
[4] L. Chen, M. Li, W. Su, M. Wu, K. Hirota, W. Pedrycz, Adaptive feature selection-based adaboost-knn with direct optimization for dynamic emotion recognition in human–robot interaction, IEEE Trans. Emerg. Topics Comput. Intell. (2019).
[5] R.W. Picard, Affective computing: from laughter to ieee, IEEE Trans. Affective Comput. 1 (1) (2010) 11–17.
[6] S. Aydın, Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips, IEEE J. Biomed. Health Inform. 24 (6) (2019) 1695–1702.
[7] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883..
[8] H. Zhang, M. Xu, Weakly supervised emotion intensity prediction for recognition of emotions in images, IEEE Trans. Multimedia (2020).
[9] P. Zhao, L. Hou, O. Wu, Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification, Knowl.-Based Syst. 193 (2020) 105443.
[10] Y. Lai, L. Zhang, D. Han, R. Zhou, G. Wang, Fine-grained emotion classification of chinese microblogs based on graph convolution networks, World Wide Web 23 (5) (2020) 2771–2787.
[11] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, H. He, Multisource transfer learning for cross-subject eeg emotion recognition, IEEE Trans. Cybern. 50 (7) (2019) 3281–3293.
[12] Y. Chen, Convolutional neural network for sentence classification Master's thesis, University of Waterloo, 2015.
[13] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: European semantic web conference, Springer, 2018, pp. 593–607.
[14] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms., in: Interspeech, 2017, pp. 1089–1093..
[15] S. Poria, N. Majumder, R. Mihalcea, E. Hovy, Emotion recognition in conversation: Research challenges, datasets, and recent advances, IEEE Access 7 (2019) 100943–100953.
[16] Z. Tang, Y. Shi, D. Wang, Y. Feng, S. Zhang, Memory visualization for gated recurrent neural networks in speech recognition, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2736–2740..
[17] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, vol. 2018, NIH Public Access, 2018, p. 2122..
[18] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2594–2604..
[19] W. Jiao, M. Lyu, I. King, Real-time emotion recognition via attention gated hierarchical memory network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8002–8009..
[20] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6818–6825..
[21] H. Zhou, J.F. Beltrán, I.L. Brito, Functions predict horizontal gene transfer and the emergence of antibiotic resistance, Sci. Adv. 7 (43) (2021) eabj5056.
[22] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations (ICLR), vol. abs/1609.02907, 2017..
[23] M. Zhang, Y. Chen, Link prediction based on graph neural networks, Adv. Neural Inf. Process. Syst. 31 (2018) 5165–5175.
[24] R. Ying, R. He, K. Chen, P. Eksombatchai, W.L. Hamilton, J. Leskovec, Graph convolutional neural networks for web-scale recommender systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 974–983..
[25] Y.-J. Choi, Y.-W. Lee, B.-G. Kim, Residual-based graph convolutional network for emotion recognition in conversation for smart internet of things, Big Data (2021).
[26] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462..
[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497..
[28] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119..
[29] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, Extraction of organic chemistry grammar from unsupervised learning of chemical reactions, Sci. Adv. 7 (15) (2021) eabe4166.
[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.
[31] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: a multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536..

Yuntao Shou is currently pursuing the undergraduate degree in the College of Computer Information and Engineering, Central South University of Forestry and Technology, Changsha, China. His research interest is object detection and emotion recognition. (e-mail: yuntaoshou@csuft.edu.cn)

Tao Meng received the Ph.D. degree the Ph.D. degree in the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include date mining, network analysis and deep learning. (Email:mengtao@hnu.edu.cn)

Wei AI received the Ph.D. degree in the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her research interests include date mining, big data, cloud computing, and parallel computing. (Email:aiwei@hnu.edu.cn)

Sihan Yang is currently pursuing the undergraduate degree in the College of Computer Information and Engineering, Central South University of Forestry and Technology, Changsha, China. His research interest is emotion recognition. (E-mail: sihanyang@csuft.edu.cn)

Keqin Li is a SUNY Distinguished Professor of computer science with the State University of New York. He is also a National Distinguished Professor with Hunan University, China. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded sys- tems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU-GPU hybrid and cooperative computing, computer ar- chitectures and systems, computer networking, machine learning, intelligent and soft com- puting. He has authored or coauthored over 830 journal articles, book chapters, and ref- ereed conference papers, and has received several best paper awards. He holds over 60 patents announced or authorized by the Chinese National Intellectual Property Administra- tion. He is among the world's top 5 most influential scientists in parallel and distributed computing based on a composite indicator of Scopus citation database. He has chaired many international conferences. He is currently an associate editor of the ACM Computing Surveys and the CCF Transactions on High Performance Computing. He has served on the editorial boards of the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Computers, the IEEE Transactions on Cloud Computing, the IEEE Transactions on Services Computing, and the IEEE Transactions on Sustainable Computing. He is an IEEE Fellow. (Email: lik@newpaltz.edu)