Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Dynamic interactive multiview memory network for emotion recognition in conversation

Jintao Wen [a,1], Dazhi Jiang [a,1], Geng Tu [a], Cheng Liu [a], Erik Cambria [b,*]

[a] Department of Computer Science, Shantou University, 515063, China
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Emotion recognition in conversation
Multimodal fusion
Dynamic interactive multiview memory network

ABSTRACT

When available, multimodal data is key for enhanced emotion recognition in conversation. Text, audio, and video in dialogues can facilitate and complement each other in analyzing speakers' emotions. However, it is very challenging to effectively fuse multimodal features to understand the detailed contextual information in conversations. In this work, we focus on dynamic interactions during the information fusion process and propose a Dynamic Interactive Multiview Memory Network (DIMMN) model to integrate interaction information for recognizing emotions. Specifically, the information fusion within DIMMN is through multiple perspectives (combining different modalities). We designed multiview layers in attention networks to enable the model to mine the crossmodal dynamic dependencies between different groups in the process of dynamic modal interaction. In order to learn the long-term dependency information, temporal convolutional networks are introduced to synthesize contextual information of a single person. Then, the gated recurrent units and memory networks are used to model the global session to detect contextual dependencies for multi-round, multi-speaker interactive emotion information. Experimental results on IEMOCAP and MELD demonstrate that DIMMN achieves better and comparable performance to the state-of-the-art methods, with an accuracy of 64.7% and 60.6%, respectively.

1. Introduction

With the booming of multimodal social media apps such as Instagram and TikTok, an increasing number of videos containing speakers' emotions are uploaded to the Internet daily. The analysis of such information has significant practical value for marketers and the brands they serve [1]. For the emotion recognition task, the speaker's emotions are often presented with different media, such as text, voice, gestures, EEG signals, etc., which can express emotions to a certain extent. Text, audio, and video (image frame) are the most commonly used data which are the fundamental forms of human communication [2–4]. Therefore, multimodal conversational emotion recognition based on videos has attracted increasing attention. There are many public multimodal datasets [5–7] and competitions [8,9] that provide a large amount of data for emotion recognition in conversation (ERC). This paper focuses on three ERC modalities: texts, audio, and videos. Much progress has been made in multimodal fusion research. Previous works used labor-intensive methods to extract features for each modality [10] manually. With the development of deep learning, deep neural network-based models have made significant progress in terms of performance.

For example, Kampman et al. [11] extracted features using convolutional neural network (CNN) for different modalities and fused them using a fully connected layer. Xu et al. [12] proposed bidirectional multilevel attention (BDLMA) model for image–text sentiment classification using complementary information from image and text data. For the ERC task, Poria et al. [13–16] used multilevel attention networks for fusing short video clips and proposed a long short-term memory (LSTM) model that was able to obtain contextual information from the corpus. Adapted dynamic memory network (DMN) [17] directly performed two-dimensional convolution of unimodal features to explore cross-modal interactions. However, existing multimodal fusion methods in ERC are conducted in a single view. Specifically, the model will give more attention to a specific elementary modality. This approach is efficient, but there is still a problem: one prominent modality will dominate the semantic information. However, the expression of emotions in ERC is mixed and diverse [18]. In addition, the emotional information from different modalities may differ (such as in Fig. 1). Therefore, it is essential to organically fuse multimodal information from multiple perspectives by semantic information mining.

[*] Corresponding author.
E-mail address: cambria@ntu.edu.sg (E. Cambria).
[1] Equal contribution.

https://doi.org/10.1016/j.inffus.2022.10.009
Received 3 November 2021; Received in revised form 7 October 2022; Accepted 9 October 2022
Available online 14 October 2022
1566-2535/© 2022 Elsevier B.V. All rights reserved.

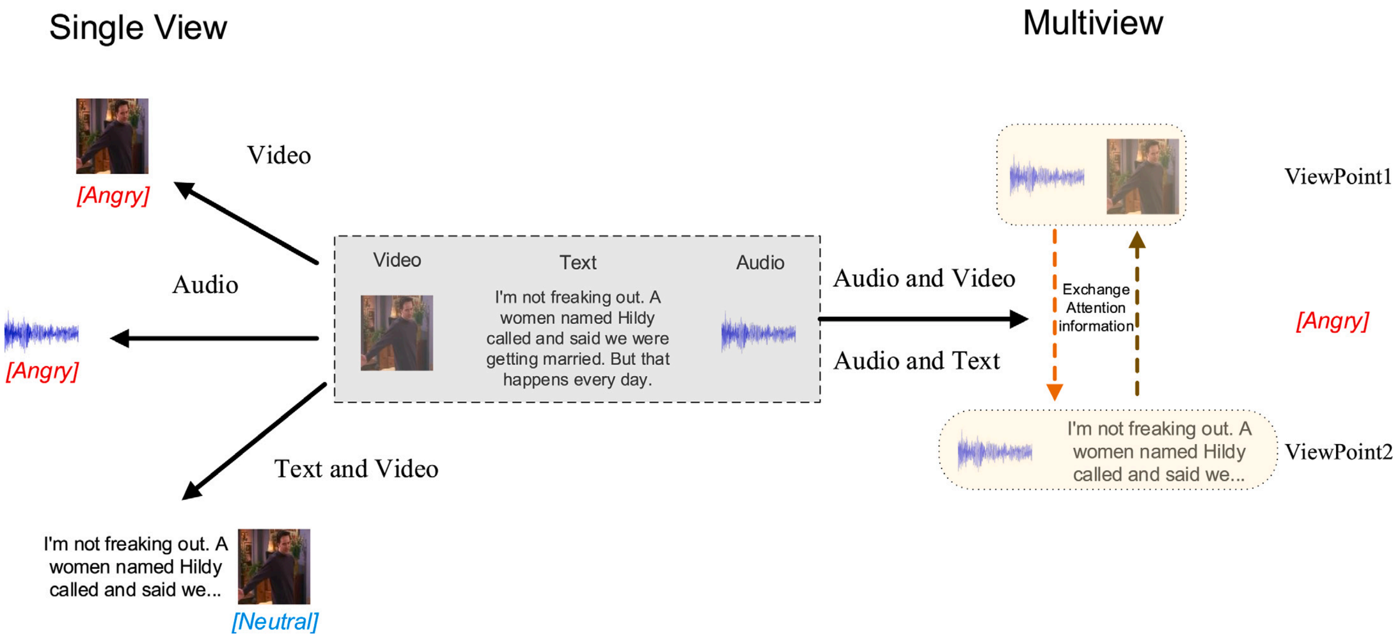
Single View
Multiview
Video
[Angry]
Audio
[Angry]
Text and Video
I'm not freaking out. A women named Hildy called and said we...
[Neutral]
Video
Text
I'm not freaking out. A women named Hildy called and said we were getting married. But that happens every day.
Audio
Audio and Video
Audio and Text
ViewPoint1
Exchange Attention information
[Angry]
I'm not freaking out. A women named Hildy called and said we...
ViewPoint2

Fig. 1. The picture compares emotion recognition results from single and multiple perspectives, respectively. They illustrate the necessity of multiview: integrating multiple views introduces more interactive information than a single view. The emotions obtained under each view may differ, so choosing an appropriate view is beneficial to model analysis.

We introduce a multiview framework to enhance the interaction during multimodal information fusion. Multimodal data are fused from different perspectives (combinations of different modalities), which not only solves the emotional inconsistency of the model from different perspectives, as shown in Fig. 1 but also dramatically enhances the robustness and accuracy of the algorithm in real scenarios. In the multimodal interactive learning module, we use two independent cross-modal multiview attention networks, which learn the emotional features of the speakers in different modalities, from text–audio and video–audio. Specifically, after each level of the attention network, the two groups of learning modules produce self-attentive fusion features and the attention-guided fusion feature. The features obtained from the fusion of multiview attention networks retain adequate information of text, audio, and video modalities simultaneously. The complementary information between the different modalities is also extracted effectively. Finally, the features of the two sets of modalities are concatenated as compact discourse representations of the speaker. In the dialogue modeling module, we believe two factors affect current emotions, different identity states (emotional contagion) and emotional context (emotional inertia). Therefore, we propose an ensemble of temporal convolutional network (TCN) and gated recurrent unit (GRU). First, each speaker is modeled independently, and TCN is used to obtain the self-emotional impact of a single person. Then, the self-emotional impact from the other speakers is modeled with the global GRU for synthesizing emotional feature fusion. Finally, with the long-term memory capability of multi-hop memory networks, context-specific information and utterance representations are combined, and the resulting information is used to classify the emotional state of the utterances.

In the theory of emotion generation and regulation [19], emotion is the summation of psychological ingredients in various ways. It will be dynamically affected by internal sensory and external situations. It shows that emotion generation is an information processing process involving information interactions and integration. The multiview framework is mainly used to solve the problem of modalities information interaction [20]. Therefore, for effectively mining the natural emotional state, the attention block based on a multiview mechanism is utilized to explore which modalities the interlocutor should use to regulate emotions. It would assign more reasonable attention weight to different modalities.

The main contributions of this paper are as follows:

1. We propose a dynamic interactive multiview memory network (DIMMN) to solve the information fusion problem in ERC tasks, especially for the inter-modal interactions during feature fusion. The multiview network fully uses the complementary information of all modalities and dynamically balances the relationship between all modalities during fusion.

2. We introduce TCN into the dialogue modeling module and build a TCN-GRU-Memory network framework. TCNs and GRUs collect the emotional impact of the speaker and other speakers, respectively, and Memory networks select different global information according to the query. The architecture can effectively fuse the information of multiperson and multi-round conversations.

3. Experimental results show that our model outperformed the existing mainstream methods on the IEMOCAP and MELD datasets. Moreover, ablation studies and comparative experiments demonstrate the effectiveness of our proposed two-stage module.

The remainder of this paper is organized as follows: Section 2 describes strengths and weaknesses of related work; Section 3 presents the interactive learning module and dialogue modeling; Section 4 carries out a series of experimental validation; finally, Section 5 provides concluding remarks.

2. Related work

The recognition of emotions has attracted the attention of various fields such as natural language processing, psychology, and cognitive science [21]. Many works have concentrated on multimodal affective computing and constructing an emotion recognition model.

Emotion recognition in conversation: Unlike single-sentence sentiment analysis, the emotion of the conversation is linked to the contextual information and speakers [22]. Considering this characteristic, most researchers used deep networks with memory functions to construct the solution model. Hazarika et al. used RNN-based deep networks for multimodal ERC. Inspired by the memory network [23], CMN [24] simulated the context information through a group of GRUs, and obtained the appropriate information from the memory network to recognize emotion. Similarly, ICON [25] further distinguished the role of the speaker in the dialogue model.

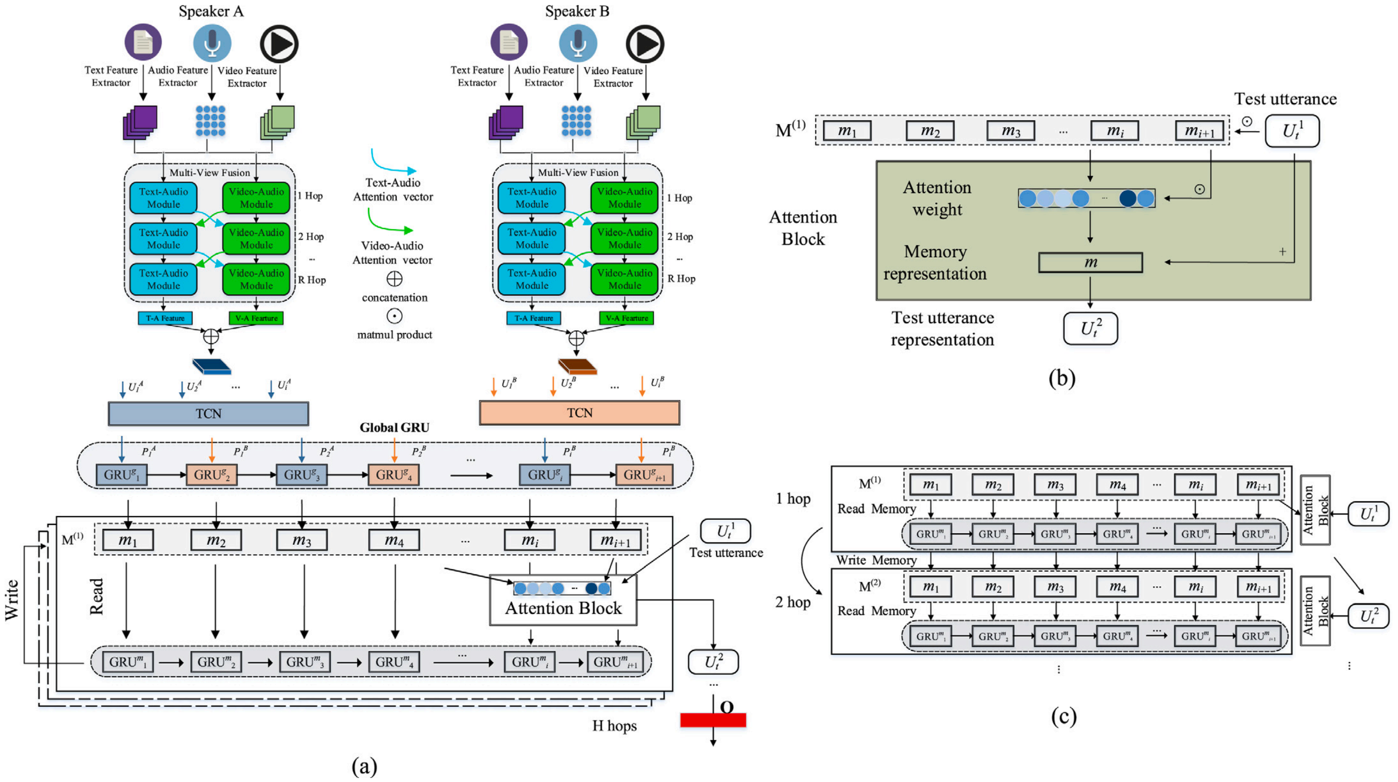

Fig. 2. (a) The framework of the attention-based interactive multiview memory network. (b) The attention block for updating the test utterance. (c) Multi-hop Write-Read scheme. The last hop of the test corpus representation $U_t^{(R+1)}$ is used to predict emotion.

Majumder et al. [26] simulated speaker states, emotion states, and global states in conversation, respectively. Zadeh et al. [15] proposed a mixed long and short-term memory based on LSTM. DialogueGCN [27] and ConGCN [28] applied graph neural networks (GCNs), which improved the context understanding. Chaturvedi et al. [29] introduced a joint model that combined CNN and fuzzy logic, which predicts the degree of a specific emotion. Li et al. [30] designed a recurrent neural network (RNN) containing different structures of generalized neural tensor blocks for emotion classification.

Recently, some novel models incorporating knowledge graph [31] and multitask learning have been proposed. COSMIC [32] obtained the speaker's internal and external states according to the causal knowledge to analyze emotion. Stappen et al. [33] explored a lexical knowledge-based extraction approach to obtain emotion understanding from video transcriptions. Zhang et al. [34] utilized reinforcement learning and domain knowledge to recognize emotion for multimodal conversational videos. Liang et al. [35] built a graph convolution network based on a dependency tree and affective commonsense to capture emotional dependencies with specific aspects. Tu et al. [36] introduced multitask graph neural network to explore subtle emotion changes. Jiang et al. [37] proposed a multitask learning framework, which can adaptively change the weight of loss per subtask. Li et al. [38] proposed a multitask learning framework based on personality characteristics and emotions. However, the above studies fuse multimodal data from a single view, and most fusion methods aim for a particular perspective, including text, audio, and video.

Multimodal fusions in affective computing: Integrating multimodal data in ERC has become a hot topic in this research. Multimodal fusion methods have gained tremendous progress in recent years. The existing fusion methods are divided into three categories [3,39], which are early fusion, late fusion, and hybrid fusion.

Early fusion [40] is also known as Feature-level fusion. It integrates features immediately after they are extracted, preparing features by computing point sums or dot products between corresponding position elements of a vector, or combining input-level features by concatenating input vectors. However, most of them depend on feature concatenation [41]. For example, Wollmer et al. [42] extracted features from each of the three modalities and then fused them at the feature level; CMF [24], ICON [25] directly concatenated multimodal features.

Late fusion (Decision-level fusion) [39,43] performs integration only after each modality outputs results (e.g., outputs classification or regression results). Late fusion first models each modality independently and then integrates the results of multiple classifiers using a fusion method such as the majority rule or weighted average. For example, Kampman et al. [11] integrated the output of each modality to make the final decision; Nojavanasghari et al. [44] trained different models for each modality and then uses a decision voting framework to combine all the models to make the final reasoning. However, late fusion also fails to explore inter-modal dynamics because the features of the various modalities cannot influence each other.

Hybrid fusion [45] is mainly performed in the middle layer of the neural network. It uses a shared layer of the network to merge neural units of multimodal data, and the intermediate layer shares parameters to fuse different modal features. Zadeh et al. [46] proposed tensor fusion networks to capture the dependency relationships within and between three modality data. Hazarika et al. [47] proposed a new framework, MISA, which projected each modality data into two different subspaces for multimodal fusion. Akhtar et al. [48] used a multitask learning approach for final sentiment and emotion classification with audio, visual, and text features.

3. Methodology

Section 3 describes the proposed dynamic interactive multiview memory network for multimodal ERC. The feature extraction processes are described for text, audio, and video in the first part of this section. Then, the model's framework is described, consisting of two modules: a multiview interactive learning module and an interactive conversation analysis module. The algorithms of both modules will be presented. Fig. 2 shows the overall architecture of the model.

3.1. Pre-extracted features

In this paper, different modalities of the interlocutor are first extracted from the video. These extracted features are fed into the conversation analysis network through the multiview fusion module to determine the conversation sentiment. We describe the feature extraction methods as follows.

3.1.1. Text features

CNN [49] is an effective method for extracting textual features from each discourse. The CNN method can learn the abstract semantic representations based on the words and n-grams of a sentence [50]. Following Kim et al. [51], we utilize a CNN that has a single convolutional layer consisting of three filters of sizes f_t^1, f_t^2, and f_t^3 (set them to 3, 4, and 5, respectively), each with a feature mapping. We use these filters for one-dimensional convolution and then perform maximum pooling on their output, and the pooled features are finally activated by rectified linear units (ReLU). The text extraction dimension is set to 100, representing the text corpus.

3.1.2. Audio features

Voice information plays a crucial role in adjusting the speakers' emotional state. The open source software openSMILE [52] is used to extract audio features. It provides High-dimension vectors for digital audio features, which contain momentous statistical notions such as formant, MFCC, MFSC, etc. In the same way, as Hazarika et al. [25], we extracted d_a audio features by openSMILE, and d_a are set to 100 dimensions.

3.1.3. Visual features

This part is a multimodal feature fusion module for the three unimodal feature representations after feature extraction. Unlike simple attention networks, the multiview attention network iterates several times interactively for each set of inputs. The result of each iteration combines the attention weight vectors of the current set itself and another set, which together lead the layer network to fuse a new feature tensor. The interaction phase can retrieve information among different modalities that were considered to be irrelevant. Each attention network includes several levels. Inspired by [53], multiview attention networks capture cross-modal interactions between two different sets of modalities.

3.2. Multiview fusion module

This part is a multimodal feature fusion module that obtains three unimodal feature representations after feature extraction, text, audio, and video. Unlike simple attention networks, the multiview attention network iterates several times interactively for each set of inputs. The result of each iteration combines the attention weight vectors of the current set itself and another set, which together lead the layer network to fuse a new feature tensor. The interaction phase can retrieve information between different modalities that were previously considered irrelevant. Each attention network includes several levels. Inspired by [53], multiview attention networks capture cross-modal interactions between two different sets of modalities.

Before feeding the feature vectors of all three modalities into the attention network, their dimensions should be unified. Let $B_{t-a} = \left[F_t, F_a\right]$ be the feature set after text–audio dimension unification, $B_{v-a} = \left[F_v, F_a\right]$ be the feature set after video–audio dimension unification, where F_t denotes text features, F_a denotes audio features, F_v denotes video features; the following experiments $B_{t-a} \in R^{d_{t-a}\times 2}, d_{t-a}$=100 and $B_{v-a} \in R^{d_{v-a}\times 2}, d_{v-a}$=100, Fig. 3 shows the process of multiview fusion. For the text–audio fusion module, the purpose is to extract the key information of text and audio. To achieve that, video features guide the fusion model to obtain the set of feature tensors.

The module calculates the attention weight vector att_{t-a} for the text–audio module by self-attention in the first-level network; In the multiview layer, the text–audio features will be weighted summed with att_{t-a} and att_{v-a} respectively to obtain the text–audio fusion tensor representation F_{t-a} and the video-guided text–audio fusion tensor representation f_{t-a}. Then, the F_{t-a} and f_{t-a} will be concatenated together and sent to the next level with a repeated process.

The video–audio fusion module is similar to the text–audio fusion. The module calculates the attention weight vector att_{v-a} and the fusion tensor representation F_{v-a} of the video–audio module by self-attention in the first level. Meanwhile, each level exchanges the attention correlation coefficients with another set to obtain the text–audio guided tensor representation f_{v-a}. Concatenate fusion tensor representation F_{v-a} and text–audio guided tensor representation f_{v-a} and feed into next level. The outcome from the last level does not need to go through the multiview layer again. The attention weight vectors and multimodal feature vectors in each level are calculated as follows:

We uniformly used the Hadamard product to calculate the attentional correlation coefficients for multiple groups of modules because it preserves sufficient information about the interactions between different modalities.

Initial state (1st level):

$$temp_{t-a}^{1} = tanh(W_{t-a} \cdot B_{t-a}), \tag{1}$$

$$temp_{v-a}^{1} = tanh(W_{v-a} \cdot B_{v-a}), \tag{2}$$

$$att_{t-a}^{1} = softmax(w_{t-a}^{T} \cdot temp_{t-a}^{1}), \tag{3}$$

$$att_{v-a}^{1} = softmax(w_{v-a}^{T} \cdot temp_{v-a}^{1}), \tag{4}$$

where $W_{t-a} \in R^{d\times d}$ represents text–audio weigh matrix, $w_{t-a} \in R^{d\times 1}$ is a column vector in W_{t-a}, $W_{v-a} \in R^{d\times d}$ represents video–audio weigh matrix, $w_{v-a} \in R^{d\times 1}$ is a column vector in W_{v-a}, $att_{t-a} \in R^{1\times 2}$ is text–audio attention weight vector, $att_{v-a} \in R^{1\times 2}$ is video–audio attention weight vector.

Next, the model will focus on B_{t-a} and B_{v-a} from two different perspectives to obtain the representations $F^{(1)}$ and $f^{(1)}$ after the fusion of the two modalities.

$$F_{t-a}^{(1)} = B_{t-a} \cdot (att_{t-a}^{1})^{T}, \tag{5}$$

$$F_{v-a}^{(1)} = B_{v-a} \cdot (att_{v-a}^{1})^{T}, \tag{6}$$

$$f_{t-a}^{(1)} = B_{t-a} \cdot (att_{v-a}^{1})^{T}, \tag{7}$$

$$f_{v-a}^{(1)} = B_{v-a} \cdot (att_{t-a}^{1})^{T}, \tag{8}$$

where $F_{t-a} \in R^{d\times 1}$ represents the result of focus on text and audio in the round, $f_{t-a} \in R^{d\times 1}$ represents the result of focus on video and audio, $F_{v-a} \in R^{d\times 1}$ represents the result of focus on video and audio in the round, $f_{v-a} \in R^{d\times 1}$ represents the result influenced by text and audio.

Starting from the ith level (i>1), the $F^{(i-1)}$ and $f^{(i-1)}$ will guide each other and calculate the new attention weight vectors to get new characterization of the ith level.

From the 2nd level to the last level (ith level, $i > 1$):

$$temp_{t-a}^{i} = tanh(W_{t-a} \cdot [F_{t-a}^{(i-1)}, f_{t-a}^{(i-1)}]), \tag{9}$$

$$temp_{v-a}^{i} = tanh(W_{v-a} \cdot [F_{v-a}^{(i-1)}, f_{v-a}^{(i-1)}]), \tag{10}$$

$$att_{t-a}^{i} = softmax(w_{t-a}^{T} \cdot temp_{t-a}^{i}), \tag{11}$$

$$att_{v-a}^{i} = softmax(w_{v-a}^{T} \cdot temp_{v-a}^{i}), \tag{12}$$

where $W_{t-a} \in R^{d\times d}$, $W_{v-a} \in R^{d\times d}$, $w_{t-a} \in R^{d\times 1}$, $w_{v-a} \in R^{d\times 1}$, $att_{t-a}^{i} \in R^{1\times 2}$, $att_{v-a}^{i} \in R^{1\times 2}$.

The steps in multiview layer are described analogous as (5)∼(8).

$$F_{t-a}^{(i)} = [F_{t-a}^{(i-1)}, f_{t-a}^{(i-1)}] \cdot (att_{t-a}^{i})^{T}, \tag{13}$$

$$F_{v-a}^{(i)} = [F_{v-a}^{(i-1)}, f_{v-a}^{(i-1)}] \cdot (att_{v-a}^{i})^{T}, \tag{14}$$

$$f_{t-a}^{(i)} = [F_{t-a}^{(i-1)}, f_{t-a}^{(i-1)}] \cdot (att_{v-a}^{i})^{T}, \tag{15}$$

$$f_{v-a}^{(i)} = [F_{v-a}^{(i-1)}, f_{v-a}^{(i-1)}] \cdot (att_{t-a}^{i})^{T}. \tag{16}$$

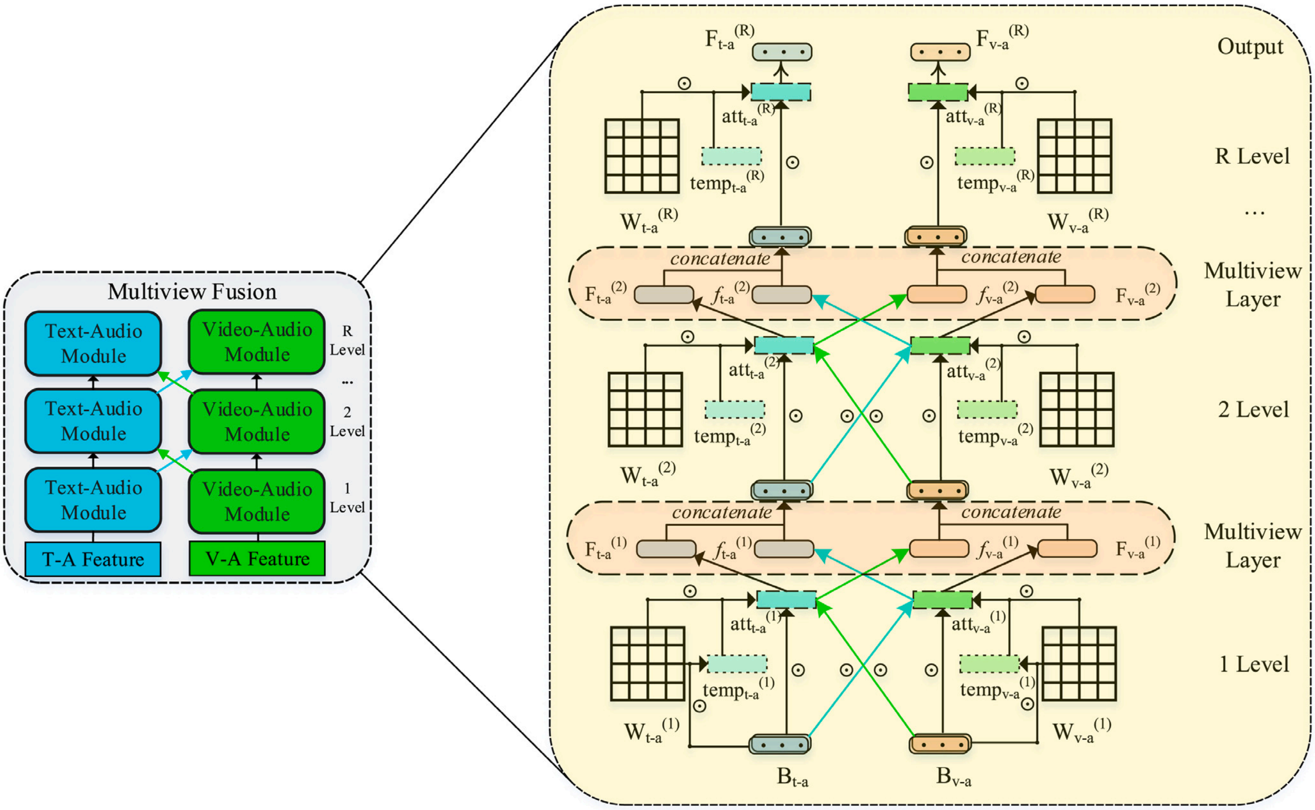

Fig. 3. The interactive learning process of the multiview attention network. It is divided into two parts: text–audio attention fusion networks and video–audio attention fusion networks, in which the two parts obtain video-guided text–audio features and text-guided video–audio features in the multiview layer, respectively, to achieve the purpose of interaction between different modalities.

The output of the last level $F_{t-a}^{(R)} \in R^{d\times 1}$ and $F_{v-a}^{(R)} \in R^{d\times 1}$ are the text–audio fusion result and video–audio fusion result, respectively. To summarize, in each multimodal fusion level, the attention weights of text–audio and video–audio are considered simultaneously. Namely, the multimodal features are fused from two different perspectives. The representations from the two perspectives are concatenated before the next round of fusion. The features learned in this way retain more critical information for the ERC task. Finally, we concatenate the outputs of the multiview attention blocks to obtain U.

$$U_X = tanh((W^f[F_{t-a}^{(R)}; F_{v-a}^{(R)}])+b^f), \tag{17}$$

$X \in \{S_A, S_B, \cdots S_X\}$, where U_X is the output corpus representation with multimodal information.

3.3. Interactive conversation module

After the fusion of internal features of a single person, we obtained a corpus representation $U_i^X = \{u_1, u_2, \dots, u_N\}, X \in \{S_A, S_B, \dots, S_X\}$ with multimodal information, where N is the number of test corpus statement. According to the DialogueRNN [26], two major impacts in the process of fusing external sentiment information of different speakers should be considered: the historical sentiment state of each speaker and the sentiment dependency relationship of the global discourse. To address these two assumptions, an independent TCN is constructed around each interlocutor to capture the single-player sentiment state. A set of GRUs is constructed around the global conversational sentiment information to capture the sentiment dependencies generated from different interlocutors through communication. After learning from these two sets of networks, the model fuses the emotional features of all speakers. Finally, a multi-hop memory network is constructed as an information base to save the sentiment state of each sentence. It is combined with specific sentences for emotion classification.

3.3.1. Independent TCN

In ERC problems, many models and methods are modeled using LSTM or GRU since RNNs can fully exploit the information's temporal and semantic information when dealing with time-series problems. However, recent works in different scenarios of time-series problems show that specific CNN structures can achieve even better results [54–59]. This paper selects TCN as a model for building individual speaker history sentiments. It is found to capture more effectively [60] after multiview feature fusion. The TCN applies a residual network and inflated convolution, which has a broader and more flexible receptive field than GRU. It means that a relatively long valid history of information can be captured (the network can further observe and predict). In the independent TCN calculation module, for each sentence of information, the input value $u_i = U_i$, TCN calculates the single person's emotional state information $p_i = TCN(u_i), i \in k$, k for each time step (the length of a sentence as a time step) denoting the size of the time window for the whole segment.

3.3.2. Global contextual GRU

Theories from cognitive science also suggest that emotional influences persist between people [61,62]. During the conversation, changes in the emotions of different speakers are transmitted to each other through their corresponding gestures, postures, and intonations. All these changes occur dynamically within the course of the conversation. While modeling the different speakers individually, this module incorporates the historical information of each speaker into the global impact to maintain a global representation of the entire context. It updates the context information iteratively at each time step in a global conversation. For an arbitrary $k \in [1, K]$, the global emotion state of the conversation is updated by taking information about the previous state s_{k-1} and the current speaker's affective state $p_k^{(x)}, x \in \{S_A, S_B, \dots, S_X\}$, which is the conversation information obtained by $u_k^{(x)}$ through the kth time step of the independent TCN computation.

This module consists of a GRU network with the following state calculation for each time step. GRU was proposed by Cho et al. [63] to capture dependencies with large time step distances in a time series. In the kth time step, its hidden state $s^k \in R^{dem}$ is computed from r^k (reset gate), up^k (update gate), the current input p^k and the previous moment state s^{k-1} as follows:

$$up^k = \sigma(V_{up}x^k + W_{up}s^{j-1} + b_{up}), \tag{18}$$

$$r^k = \sigma(V_r x^k + W_r s^{j-1} + b_r), \tag{19}$$

$$v^k = tanh(V_h x^k + W_h(s^{k-1} \otimes r^j) + b_h), \tag{20}$$

$$s^k = (1 - up^k) \otimes v^k + up^k \otimes s^{k-1}. \tag{21}$$

In this work, the global GRU with input $x^k = p_X^k, X \in \{S_A, S_B, \dots, S_X\}$

If $X = \mathrm{S_A}$

$$s^k = GRU^g(s^{k-1}, p_A^k). \tag{22}$$

Else if $X = \mathrm{S_B}$

$$s^k = GRU^g(s^{k-1}, p_B^k). \tag{23}$$

3.3.3. Conversation memory network

After the overall operation of the global context module, the GRU^g network produces a series of memories $M = [s_1, \dots, s_k] \in R^{dem\times k}, dem$ as the embedding dimension of each sentence after modal fusion. These memories M contain dynamic influences from each time window in history (K in total). They act as a repository of contextual memories from which person-specific data can be selectively fitted into the testing corpus for obtaining distinguishable features. To this end, a set of GRU^g networks are used to perform each hop of the read or write cycle task separately, where the test utterance u_t is combined with soft attention during the read or write to refine it into a context-aware representation.

This module contains the information of all conversations, so it is insufficient to use RNN-based networks because their memory ability is so finite that they cannot store the global memory. TCNs can retain the whole historical information, but their space overhead is so large that we can hardly do it under limited conditions. Memory networks can solve both problems. The need for "multi-hop" is inspired by the related work on memory networks [41,43,64], which shows the importance of multiple read or write iterations for performing transitive reasoning. There is a risk of missing the underlying memory with just one cycle; however, multiple iterations will help improve the attention head's concentration. Fig. 2(b) and (c) show the attention block and multi-hop read/write process. At the rth hop, the following results are calculated.

Read Memory Unit: We read memories from the memory unit $M^{(r)}$ in the rth iteration utilizing an attention mechanism. Firstly, each memory $m_k^r \in M^{(r)}$ applies matmul product computation with the test utterance $u_t^{(r)}(M^{(1)} = M$ at the first iteration).

The results were normalized to produce an attention-based weight vector $att^{(r)} \in \mathbb{R}^K$, which kth normalized score manifests the correlation of the kth unit to the test corpus. The attention vector $att^{(r)}$ was computed as follows:

$$att^{(r)} = softmax((M^{(r)})^T \cdot u_t^{(r)}). \tag{24}$$

The attention weight vector $att^{(r)}$ is weight-summed with each sentence and used to find a new weighted memory representation:

$$m^{(r)} = \sum_{k=1}^{K} att_k^{(r)} \cdot m_k^{(r)} = M^{(r)} \cdot att^{(r)}, \tag{25}$$

$m^{(r)}$ as a weighted memory, which is both a summary of contextual emotion and information based on the test corpus. Finally, the weighted memory $m^{(r)}$ is used to update test utterance $u_t^{(r)}$ as follows:

$$u_t^{(r+1)} = tanh(m^{(r)} + u_t^{(r)}). \tag{26}$$

Table 1
The split of datasets.

Dataset	Dialogues			Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1039	114	280	9989	1109	2610

Write Memory Unit: After each hop of the read operation, the memory will be updated in the next hop. We use several GRU units as a network, called GRU^m, that takes as input the memory unit $M^{(r)}$ of the rth iteration and reprocesses this sequence through the GRU network to generate a new memory $M^{(r+1)}$ for the $(r+1)$th hop, i.e., $M^{(r+1)} = GRU^m(M^{(r)})$. During the whole hops, writing memory can be regarded as a stacked RNN in which each hop improves the representation of the RNN output. The GRU^ms' parameters are shared.

Predication: For the final classification result, the dimension of vector o is the number of emotion classes C, namely $o \in R^C$. Cross-entropy is used as a loss function for training.

$$o = softmax(W^o u_t^{(R+1)} + b). \tag{27}$$

For the final classification result, the dimension of vector o is the amounts of emotion classes C, namely $o \in R^C$. Cross-entropy is a loss function to reflect the cost metric of training.

4. Experiment

In this section, our work conducts comprehensive experiments to present the experimental results on different network structures compared with various baselines. We also provide necessary analyses and discussions for further argumentation to validate our proposed model.

4.1. Datasets detail

IEMOCAP: IEMOCAP [5] is the most widely used multimodal dataset in ERC that contains ten two-person dialogues between speakers, including video, speech, facial motion capture, and text transcription. The IEMOCAP database is divided into five parts, each pair being assigned to a different conversation scene. The dataset considers six emotional types for the classification task: anger, happiness, sadness, neutrality, excitement, and frustration. We extract the three modalities and assign the dialogue ID to ensure they are aligned. After pre-extraction, we use 80% of the utterances in the training and verification set and the remaining 20% in the test set. The multimodal dialogue split scheme is similar to Hazarika et al. [25] and Majumder et al. [26].[2]

MELD: MELD [6] is created by enhancing and extending the sentiment line dataset. It consists of 1433 dialogues selected from the American drama "Old Friends" with 13,708 sentences, including video, text, speech, and other data content. MELD labels each discourse in the dialogues with seven emotions: anger, disgust, sadness, happiness, neutrality, surprise, and fear. Table 1 shows our division of the dataset.

4.2. Baselines

DIMMN is compared with various state-of-the-art corpus-level emotion classification models for multimodal.

c-LSTM [14] A hierarchical structure is designed to classify the corpus using the neighboring corpus (of the same speaker) as the context.

CMN [24] The model uses GRU in ERC to extract each speaker's conversation information separately, then feeds their history into the memory network and makes the final prediction.

[2] https://github.com/declare-lab/conv-emotion

Table 2
Different models on the IEMOCAP dataset.

Models	Happy		Sad		Neutral		Angry		Excited		Frustrated		Avg	
	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$
c-LSTM	25.7	35.6	58.6	69.3	56.2	53.5	70.0	66.3	58.8	61.1	67.4	62.4	57.6	59.0
CMN	25.0	30.4	55.9	62.4	53.2	56.2	67.6	64.6	69.9	67.9	71.7	63.1	59.9	59.4
MFN	24.0	34.1	65.6	70.2	55.5	52.1	**72.3**	66.8	64.3	62.1	67.9	62.5	60.1	59.9
ICON	23.6	32.8	70.6	74.4	59.9	60.6	68.2	**68.2**	72.2	68.4	71.9	66.2	62.8	62.9
Dialogue-RNN	25.7	33.2	**75.1**	**78.8**	58.6	59.2	64.7	65.3	80.3	71.9	61.2	58.9	63.4	62.8
QMNN	**41.3**	39.7	72.9	68.3	54.1	55.3	65.4	62.6	66.0	66.7	55.6	62.2	60.8	59.9
FERNet	38.9	**40.1**	72.7	70.2	**67.2**	**61.5**	66.5	62.4	68.9	68.2	50.4	58.6	61.8	61.6
TL-ERC	–	–	–	–	–	–	–	–	–	–	–	–	–	58.5
Our Modelremove multiview	22.9	30.8	69.8	72.3	60.9	59.5	65.3	60.9	67.9	66.7	70.9	64.4	63.1	62.2
Our Modelremove TCN	25.0	29.3	62.9	72.8	55.5	57.9	62.9	62.8	77.6	70.3	**76.1**	65.3	63.6	63.0
Our Model	24.3	30.2	64.5	74.2	57.3	59.0	61.8	62.7	**81.3**	**72.5**	75.9	**66.6**	**64.7**	**64.1**

Table 3
Different models on the MELD dataset.

Models	Anger		Disgust		Fear		Joy		Neutral		Sadness		Surprise		Avg	
	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$	acc	$w-F1$
test-CNN	34.5	-	8.2	-	3.7	-	49.4	-	74.9	-	21.0	-	45.5	-	55.0	-
c-LSTM	43.4	46.0	**23.7**	**28.6**	9.4	5.4	**54.5**	**58.7**	76.7	73.8	24.3	**33.0**	51.0	46.7	57.6	55.9
Dialogue-RNN	43.7	41.5	7.9	1.7	**11.7**	7.2	54.4	50.7	77.4	73.5	**34.6**	23.8	**52.5**	**49.4**	60.3	57.4
QMNN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**60.8**	58.0
NTN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.2	-
Our Model	**53.0**	**47.6**	7.6	6.7	9.7	**8.8**	53.2	52.3	**82.3**	**76.0**	3.8	4.6	47.3	46.9	60.6	**58.6**

MFN [65] A multimodal fusion method based on the outer tensor product (Outer product) is proposed to explore the unimodal and multimodal interactions in network models using multiview learning.

ICON [25] The model adds a GRU network to the CMN to establish a global dialogue and realize the interaction of the two-person dialogue process.

DialogueRNN [26] The influence of the different roles (speaker and listener) in the conversation on the sentiment is proposed. The model consists of two sets of GRUs for tracking the speaker's state and the context during the conversation.

QMNN [66] Multimodal fusion and dialogue modeling is achieved through a quantum-inspired neural network, which is a new perspective in the ERC.

FERNet [67] It is a fine-grained extraction and reasoning network based on two components for generating target-specific historical discourse expressions.

TL-ERC [68] Transfer learning was introduced into ERC. The parameters of the pre-trained model are transferred to another emotion model.

NTN [69] A neural tensor network consisting of bilinear, linear, and bias terms is used for ERC. In addition, they propose a factorization method that makes the neural tensor network more efficient.

4.3 Results and discussion

To validate the effectiveness of our proposed approach, we first compared our model with the advanced multimodal-based modeling approach described above on the IEMOCAP dataset. Python 3.6, TensorFlow 1.13, and Keras libraries are used on an NVIDIA RTX 3060 GPU for our experiments. The running period of each epoch in this environment is about 30 s, which is relatively efficient. The final experimental results are presented in Table 2. Two metrics, $average-acc$ and $F1$ score, are used as the basis for evaluation. Our model achieves an average accuracy of 64.7%, and the results are the best here. It is important to note that $average-acc$ here refers to a weighted average rather than simply averaging over the categories.

Contrasted with the advanced architecture like DialogueRNN, QMNN, and FERNet, our model achieves a 1.3% advance in weighted average accuracy and 1.2% improvement in the $F1$ score, respectively. For all 14 evaluation metrics of the IEMOCAP dataset, our model outperforms other methods in 5 metrics and achieves competitive performance in the remaining metrics. Ablation experiments are also included to validate the effectiveness of multiview fusion and TCN. We control variables for the modal fusion part and TCN, respectively. The experimental results show that multiview fusion makes more effective use of multimodal information, which guides features of different modalities to make correct judgments according to their "strengths". Meanwhile, using TCNs instead of GRUs to model contextual information of a single speaker has some improvement in the results. Dilatation convolution plays an important role in them. Here we are using exponential dilation convolution. The dilation factor is related to the number of layers and is set to 2^i. We artificially "force" the network to use only the historical step information within the range of dilation factor. However, the RNN network can only input all the historical information by default and cannot achieve fine control of the input information. It is the main reason to improve the result in our opinion. In general, DIMMN can effectively capture cross-modal information from multimodal information with cross-modal interactions and have a strong ability to capture contextual sentiment, both of which would improve the performance of the ERC task. Fig. 4, and Fig. 5 show the confusion matrix under the IEMOCAP dataset.

Next, our model is evaluated on the MELD dataset in the same environment, the largest multiparty and multimodal ERC dataset. We used the textual audio modal per the MELD setup to compare our model with state-of-the-art methods in a multimodal setting. Unlike IEMOCAP, two sets of textual and audio attentive modules are utilized respectively in the multiview interactive learning phase due to the lack of video modal. The experimental results are summarized in Table 3. It can be seen that our model performs best in the $F1$ score. Figs. 6 and 7 show the confusion matrix for the MELD dataset.

During the experiment, we explored the effects of different modal groupings of interactive multiview memory networks on the experimental results. We set up four groups of fusion schemes to be compared on the IEMOCAP dataset: text, audio, and video connection; text–audio and video–audio interactive learning; text–video and audio–video interactive learning; and video–text and audio–text interactive learning.

Table 4
Controlled experiment of different H on IEMOCAP.

Models	Happy		Sad		Neutral		Angry		Excited		Frustrated		Avg	
	acc	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*
H=2	23.6	**32.2**	64.5	63.8	**58.6**	**59.3**	61.6	62.2	76.2	70.7	75.2	58.3	63.2	62.7
H=3	24.3	30.2	64.5	**74.2**	57.3	59.0	61.8	62.7	**81.3**	**72.5**	**75.9**	**66.6**	**64.7**	**64.1**
H=4	**25.0**	31.9	**64.9**	74.0	57.3	59.0	61.8	62.7	77.9	71.9	**75.9**	**66.6**	64.2	63.7
H=5	23.8	30.9	64.2	74.0	57.1	58.7	**64.7**	**65.9**	76.0	70.6	75.6	60.0	63.8	62.7

Table 5
Controlled experiment of different K on IEMOCAP.

Models	Happy		Sad		Neutral		Angry		Excited		Frustrated		Avg	
	acc	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*
K=3	24.0	30.0	63.8	73.2	56.7	58.0	60.0	60.9	76.9	69.8	74.5	65.8	63.5	62.8
K=4	24.1	30.1	**64.6**	74.1	57.1	58.8	61.3	62.0	79.6	71.0	75.1	66.4	64.3	64.0
K=6	**24.3**	**30.2**	64.5	**74.2**	**57.3**	**59.0**	**61.8**	**62.7**	**81.3**	**72.5**	**75.9**	66.6	**64.7**	**64.1**
K=8	**24.3**	**30.2**	64.5	73.8	**57.3**	58.7	**61.8**	62.4	80.3	71.6	**75.9**	**66.8**	64.6	**64.1**

Table 6
Controlled experiment of different L on IEMOCAP.

Models	Happy		Sad		Neutral		Angry		Excited		Frustrated		Avg	
	acc	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*	*acc*	*w − F1*
L=5	22.2	27.3	60.6	69.2	56.8	52.1	58.8	60.7	78.6	70.8	74.8	64.6	63.7	62.8
L=6	**24.3**	**30.2**	**64.5**	**74.2**	**57.3**	**59.0**	61.8	62.7	**81.3**	**72.5**	**75.9**	**66.6**	**64.7**	**64.1**
L=7	**24.3**	30.0	63.7	71.6	56.8	54.2	**62.4**	**63.5**	79.9	71.7	74.8	65.2	64.5	63.9
L=8	23.6	28.8	64.1	74.0	56.8	53.6	61.2	61.3	77.9	70.9	75.3	66.0	64.4	63.6

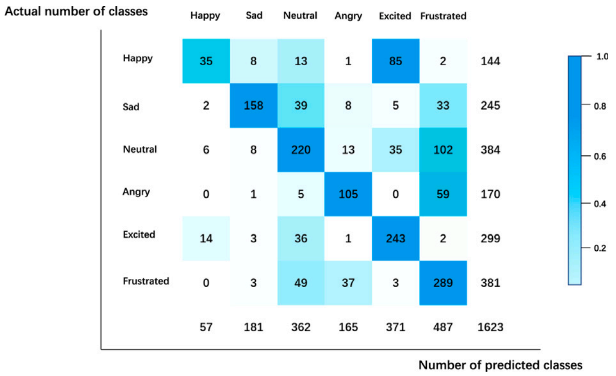

Fig. 4. Confusion matrix under recall metrics of IEMOCAP dataset.

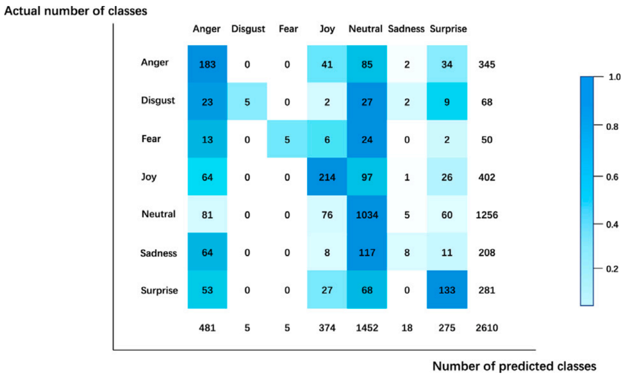

Fig. 6. Confusion matrix under recall metrics of MELD dataset.

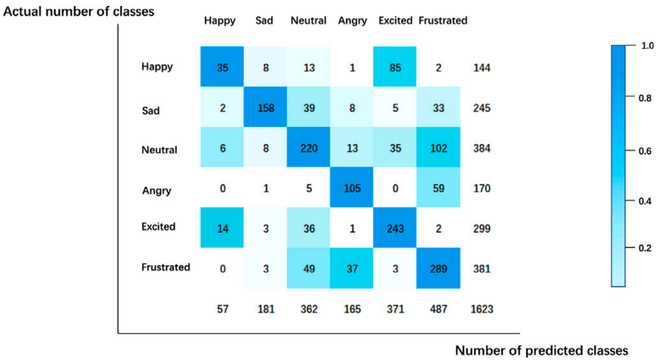

Fig. 5. Confusion matrix under precision metrics of IEMOCAP dataset.

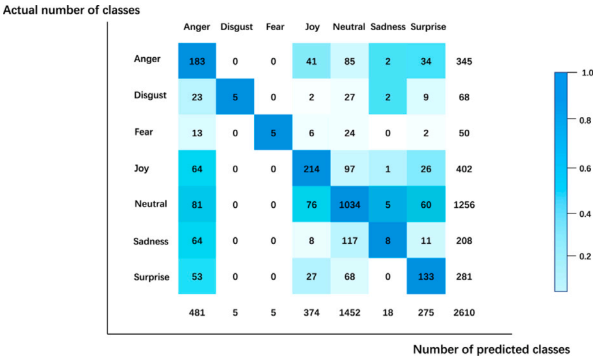

Fig. 7. Confusion matrix under precision metrics of MELD dataset.

It can be found from the experiments that the text–audio and video–audio have the best results. This result indicates that the dialogue information is practical after the multiview memory network. Both video–text and audio–text views retain too much text information while ignoring the information from other modalities.

The resulting features are not conducive to the next stage of sentiment analysis. Table 7 shows the experimental results of different fusion approaches. We explored the effects of the level numbers H and the convolutional kernel size K and layer numbers L of the TCN in DIMMN on the experimental results. Orthogonal and control experiments are designed to select the optimal parameters, respectively.

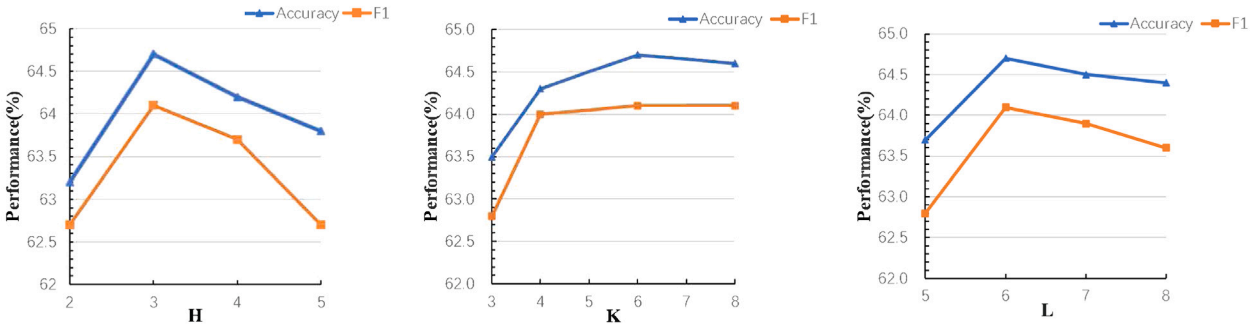

Fig. 8. Parameter analysis results of DIMMN.

Table 7
Different modal combinations on IEMOCAP.

Modal	acc	$w-F1$
A+V+T	62.8	62.1
TV+AV	61.9	60.9
TA+VA	**64.7**	**64.1**
VT+AT	61.0	59.5

Table 8
Orthogonal tests on the IEMOCAP dataset.

IDs	H	K	L	Result(acc)
1	2	3	5	62.8
2	2	4	6	63.1
3	2	6	7	63.2
4	2	8	8	63.1
5	3	3	6	63.5
6	3	4	5	63.8
7	3	6	8	**64.5**
8	3	8	7	64.2
9	4	3	7	63.6
10	4	4	8	63.8
11	4	6	5	64.0
12	4	8	6	63.9
13	5	3	8	63.2
14	5	4	7	63.5
15	5	6	6	63.8
16	5	8	5	63.7

The orthogonal test is an efficient experimental method [70]. It simultaneously considers three factors, H, K, and L, and some representative combinations are selected for the test based on orthogonality. These representative combinations conform to a uniform distribution, and $L_{(16)}(3^4)$ is set to represent the orthogonal test 16 times. The orthogonal experiment showed us the highest results of 64.5% for H=3, K=6, and L=8. After the orthogonal experiments, the number of attention network cycles is set as H, the convolutional kernel size as K, and the number of layers as L of TCN. We focus on the respective effects with H=3, K=6, and L=8 as the control variables. The final experimental results from Tables 2 and 3 show that our model achieves the best performance when H=3, K=6, and L=6. The performance of our proposed method becomes better as the number of cycles increases. It means that multiple shifted attention interactions can effectively capture more cross-modal interactions, and the TCN can adequately capture single-person emotions from the features. The classification performance becomes worse when H is greater than three due to the redundant information introduced in the feature fusion process. As the kernel_size and depth of the TCN network increase (until K=6, L=6), the classification results also improve significantly. It indicates that increasing the receptive field is more effective in extracting sentiment information from the features after modal fusion, but increasing the number of network layers and the size of kernel_size will lead to overfitting.

The number of model parameters and the computational time complexity also increases. Finally, H, K, and L are set to 3, 6, and 6, respectively, to obtain the best performance. Fig. 8 and Table 8 show the results of the orthogonal tests, Table 4, Table 5, and Table 6 show the results of the control experiments with different parameters, respectively.

5 Conclusion

This paper introduced a dynamic interactive multiview memory network for emotion recognition in multimodal conversations, which utilizes a multigroup attention network to learn features interactively across modalities. Our model captures cross-modal interactions from multimodal information and fuses the global information of multiple interlocutors through an ensemble of temporal convolutional networks and gated recurrent units. Both modules significantly improve the accuracy of multimodal dialogue sentiment analysis by obtaining more discriminative discourse features. Empirical evaluation of the public multimodal dialogue dataset shows that our approach achieves comparable performance with the state-of-the-art methods. As future work, we plan to introduce more modalities in interactive multiview learning and refine the images of different scenes, in order to further improve emotion recognition in multimodal conversations.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore under its AME Programmatic Funding Scheme (Project #A18A2b0046). The research is also funded by the National Natural Science Foundation of China (62106136, 61902231), Natural Science Foundation of Guangdong Province, China (2019A1515010943), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province, China (Special Projects in Artificial Intelligence) (2019KZDZX1030), 2020 Li Ka Shing Foundation, Hong Kong Cross-Disciplinary Research Grant (2020LKSFG04D), and Science and Technology Major Project of Guangdong Province, China (STKJ2021005).

References

[1] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: LREC, 2022, pp. 3829–3839.

[2] N. Sebe, H. Aghajan, T. Huang, N. Magnenat-Thalmann, C. Shan, Special issue on multimodal affective interaction, IEEE Trans. Multimed. 12 (6) (2010) 477–480.

[3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.

[4] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics, in: IEEE SSCI, Singapore, 2013, pp. 108–117.

[5] 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.

[6] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: ACL, 2019, pp. 527–536.

[7] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Trans. Affect. Comput. 3 (1) (2011) 5–17.

[8] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 3–10.

[9] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, B.W. Schuller, The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress, in: Proceedings of ACM Multimedia, 2021, pp. 5706–5707.

[10] F. Wu, Y. Huang, Y. Song, S. Liu, Towards building a high-quality microblog-specific Chinese sentiment lexicon, Decis. Support Syst. 87 (2016) 39–49.

[11] O. Kampman, E.J. Barezi, D. Bertero, P. Fung, Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 606–611.

[12] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, Y. He, Visual-textual sentiment classification with bi-directional multi-level attention networks, Knowl.-Based Syst. 178 (2019) 61–73.

[13] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2539–2544.

[14] 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.

[15] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: AAAI, 2018, pp. 5642–5649.

[16] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: ICDM, 2017, pp. 1033–1038.

[17] S. Xing, S. Mai, H. Hu, Adapted dynamic memory network for emotion recognition in conversation, IEEE Trans. Affect. Comput. (2020).

[18] Y. Ma, K.L. Nguyen, F. Xing, E. Cambria, A survey on empathetic dialogue systems, Inf. Fusion 64 (2020) 50–70.

[19] J.J. Gross, L. Feldman Barrett, Emotion generation and emotion regulation: One or two depends on your point of view, Emotion Rev. 3 (1) (2011) 8–16.

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

[21] R.W. Picard, Affective computing: from laughter to IEEE, IEEE Trans. Affect. Comput. 1 (1) (2010) 11–17.

[22] Y. Susanto, A. Livingstone, B.C. Ng, E. Cambria, The hourglass model revisited, IEEE Intell. Syst. 35 (5) (2020) 96–102.

[23] J. Weston, S. Chopra, A. Bordes, MEMORY NETWORKS, 2014, arXiv preprint arXiv:1410.3916.

[24] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: NAACL, 2018, pp. 2122–2132.

[25] 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.

[26] 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, 2019, pp. 6818–6825.

[27] 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.

[28] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations., in: IJCAI, 2019, pp. 5415–5421.

[29] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.

[30] W. Li, W. Shao, S. Ji, E. Cambria, Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.

[31] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: Representation, acquisition and applications, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2022) 494–514.

[32] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, COSMIC: COmmonSense knowledge for emotion identification in conversations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 2470–2481.

[33] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95.

[34] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1034–1047.

[35] B. Liang, H. Su, L. Gui, E. Cambria, R. Xu, Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks, Knowl.-Based Syst. 235 (2022) 107643.

[36] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, D. Jiang, Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models, Knowl.-Based Syst. (2021).

[37] D. Jiang, R. Wei, H. Liu, J. Wen, G. Tu, L. Zheng, E. Cambria, A multitask learning framework for multimodal sentiment analysis, in: 2021 International Conference on Data Mining Workshops, ICDMW, IEEE, 2021, pp. 151–157.

[38] Y. Li, A. Kazemeini, Y. Mehta, E. Cambria, Multitask learning for emotion and personality traits detection, Neurocomputing 493 (2022) 340–350.

[39] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.

[40] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 973–982.

[41] Y. Gu, X. Li, S. Chen, J. Zhang, I. Marsic, Speech intention classification with multimodal deep learning, in: Canadian Conference on Artificial Intelligence, Springer, 2017, pp. 260–271.

[42] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.

[43] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, arXiv preprint arXiv:1606.06259.

[44] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, L.-P. Morency, Deep multimodal fusion for persuasiveness prediction, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 284–288.

[45] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133.

[46] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[47] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[48] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, S. Kurohashi, All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework, IEEE Trans. Affect. Comput. (2019).

[49] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.

[50] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 655–665, http://dx.doi.org/10.3115/v1/P14-1062, URL https://aclanthology.org/P14-1062.

[51] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, http://dx.doi.org/10.3115/v1/D14-1181, URL https://aclanthology.org/D14-1181.

[52] 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.

[53] X. Yang, S. Feng, D. Wang, Y. Zhang, Image-text multimodal emotion classification via multi-view attentional network, IEEE Trans. Multimed. (2020).

[54] S. Bai, J.Z. Kolter, V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, Universal Language Model Fine-Tuning for Text Classification, 2018.

[55] C. Lea, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks: A unified approach to action segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 47–54.

[56] I. Kotenko, I. Saenko, S. Ageev, Hierarchical fuzzy situational networks for online decision-making: Application to telecommunication systems, Knowl.-Based Syst. 185 (2019) 104935.

[57] J. Ren, W. Wu, G. Liu, Z. Chen, R. Wang, Bidirectional gated temporal convolution with attention for text classification, Neurocomputing 455 (2021) 265–273.

[58] G. Guo, W. Yuan, Short-term traffic speed forecasting based on graph attention temporal convolutional networks, Neurocomputing 410 (2020) 387–393.

[59] P. Liu, J. Liu, K. Wu, CNN-FCM: System modeling promotes stability of deep learning in time series prediction, Knowl.-Based Syst. 203 (2020) 106081.

[60] G. Xiao, G. Tu, L. Zheng, T. Zhou, X. Li, S.H. Ahmed, D. Jiang, Multimodality sentiment analysis in social internet of things based on hierarchical attentions and CSAT-TCN with MBM network, IEEE Internet Things J. 8 (16) (2021) 12748–12757, http://dx.doi.org/10.1109/JIOT.2020.3015381.

[61] T.L. Chartrand, J.A. Bargh, The chameleon effect: the perception–behavior link and social interaction., J. Personal. Soc. Psychol. 76 (6) (1999) 893.

[62] C. Navarretta, Mirroring facial expressions and emotions in dyadic conversations, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 469–474.

[63] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734, http://dx.doi.org/10.3115/v1/D14-1179, URL https://aclanthology.org/D14-1179.

[64] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: International Conference on Machine Learning, PMLR, 2016, pp. 1378–1387.

[65] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: AAAI, 2018, pp. 5634–5641.

[66] Q. Li, D. Gkoumas, A. Sordoni, J.-Y. Nie, M. Melucci, Quantum-inspired neural network for conversational emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 13270–13278.

[67] Y. Guo, Z. Wu, M. Xu, FERNet: Fine-grained extraction and reasoning network for emotion recognition in dialogues, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 37–43.

[68] D. Hazarika, S. Poria, R. Zimmermann, R. Mihalcea, Conversational transfer learning for emotion recognition, Inf. Fusion 65 (2021) 1–12.

[69] W. Li, L. Zhu, E. Cambria, Taylor's theorem: A new perspective for neural tensor networks, Knowl.-Based Syst. 228 (2021) 107258.

[70] K.M. Sallam, S.M. Elsayed, R.A. Sarker, D.L. Essam, Multi-method based orthogonal experimental design algorithm for solving CEC2017 competition problems, in: 2017 IEEE Congress on Evolutionary Computation, CEC, IEEE, 2017, pp. 1350–1357.