Digital Object Identifier 10.1109/ACCESS.2020.3005664

Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations

HELANG LAI1,2, HONGYING CHEN1, AND SHUANGYAN WU1
1School of Computer Science, South China Normal University, Guangzhou 510631, China
2Guangdong Justice Police Vocational College, Guangzhou 510520, China
Corresponding author: Hongying Chen (chenhy01@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572234, in part by the 13th Five Year Plan of Guangdong Institute of Higher Education Research on Higher Education of Young Teachers in Colleges and Universities, in 2019, under Grant 19GGZ070, and in part by the Guangdong Science and Technology Collaborative Innovation Center for Judicial Administration, China, under Grant 2019B110210002.

ABSTRACT Multimodal emotion detection (MED) in interactive conversations is extremely important for improving the overall human-computer interaction experience. Present research methods in this domain do not explicitly distinguish the contexts of a test utterance in a meaningful way while classifying emotions in conversations. In this paper, we propose a model, named different contextual window sizes based recurrent neural networks (DCWS-RNNs), to differentiate the contexts. The model has four recurrent neural networks (RNNs) that use different contextual window sizes. These window sizes can represent the implicit weights of different aspects of contexts. Further, four RNNs are independently to model the different aspects of contexts into memories. Such memories with the test utterance are then merged using attention-based multiple hops. Experiments show DCWS-RNNs outperforms the compared methods on both the IEMOCAP and AVEC datasets. Case studies on the IEMOCAP dataset also demonstrate that our model has excellent performance to capture the emotional dependent utterance that is most relevant to the test utterance and assigned to the highest attention score.

INDEX TERMS Interactive conversations, contextual window sizes, emotion detection, multimodal, recurrent neural network.

I. INTRODUCTION Emotions are part of human nature and play an important role in our daily life. The development of machines that can detect and comprehend emotions has been a long-standing goal of artificial intelligence (AI). With the growth of deep learning [1] and neural network [2], many researchers have been paying increasing attention to multimodal emotion detection (MED). This can help in numerous fields like counseling, public opinion mining, financial forecasting, intelligent systems, dialogue systems, human-computer interaction, and so on. In all these applications, creating empathetic agents is an urgent need as it can improve human-computer interaction experience in conversations. Although previous significant research work has been carried out on MED using audio, visual, and text modalities [3], [4], there is little work in interactive conversations. According to [5], [6], modeling contexts and analyzing emotional dynamics [7] in conversations pose complex challenges. This is because of the presence of intricate dependencies between the affective states of speakers participating in the conversation [8]. In this paper, we propose a method, named different contextual window sizes based recurrent neural networks (DCWS-RNNs), to address the issues of MED in utterance level (generally, a unit of speech bounded by breathes or pauses of the speaker).

In interactive conversations, self and inter-speaker emotional influence are two vital properties exited in emotional dynamics [9], [10]. Self-emotional influence also termed as emotional inertia, relates to the degree to which a person’s feelings carry over from one moment to another [11], or in other words, copes with the aspect of emotional influence that speakers have on themselves during the conversation [12]. On the other hand, inter-speaker emotional influence is
DCWS-RNNs utilizes four recurrent neural networks (RNNs) to capture the most emotional dependent utterance of utterances of each aspect is its contextual window size. The number of aspects of the contexts of a test utterance in the interactive conversations should divide into four aspects, and the implicit weights of these different aspects can be represented using different contextual window sizes. Each aspect is part of contexts of the test utterance. To our knowledge, it is another property that the counterparts induce into a speaker. That means speakers may follow the emotions of their counterparts during the course of a conversation [13]. Figure 1 demonstrates a short conversation extracted from the IEMOCAP dataset, which is involving these two properties of both self and inter-speaker emotional influence.

Many studies in MED have been done by exploiting contextual information present in the conversation history [8], [13], [14]. However, they do not propose a meaningful way to distinguish the contexts of a test utterance in interactive conversations. In other words, they do not pay attention to the implicit weights of different aspects of contexts, just only using simple or equal treatments and leading to capturing emotional dynamics less effectively. In contrast, we develop a model that is to distinguish the contexts. Further, our model represents the implicit weights of different aspects of contexts using different contextual window sizes. The model is based on such an assumption: the contexts of a test utterance in the interactive conversations should divide into four aspects, and the implicit weights of these different aspects can be represented using different contextual window sizes. Each aspect is part of contexts of the test utterance. Although these four different aspects of contexts are not necessarily independent, their separate modeling exhibit superior performance in comparison with the state-of-the-art in the experiments, which is the evidence to prove that our hypothesis is effective.

The DCWS-RNNs first represents all the utterances in a video with multimodal (text, audio, and visual) features. Next, for classifying a test utterance (i.e. $u_t$), four different aspects of the contexts of $u_t$ are respectively generated by collecting corresponding utterances. The number of utterances of each aspect is its contextual window size. To capture the most emotional dependent utterance of $u_t$, DCWS-RNNs utilizes four recurrent neural networks (RNNs) to independently process these four different aspects. These RNNs are based on gated recurrent units (GRUs). One RNN is assigned to one aspect. The number of GRUs of each RNN is equal to the number of utterances of its assigned aspect. Each RNN models its corresponding aspect into memory cells using GRUs. These memory cells are then performed an attention mechanism, producing a weighted representation. This representation denotes the summary of the memories. All weighted representations of four aspects are then merged with $u_t$ using an addition operation. This is done to improve the representation of the test utterance through modeling the emotional dynamics from different aspects of contexts. The whole operation process is repeated based on a multiple hop scheme. Finally, the emotion representation of $u_t$ is obtained after $H$ hops, which is used to classify its emotion category. The contributions of this paper can be summarized as follows:

- We propose a DCWS-RNNs model to distinguish the contexts of the test utterance. To our knowledge, it is the first time to represent the implicit weights of these different aspects of contexts using different contextual window sizes.
- We adopt a multimodal approach to represent the utterances. This can provide comprehensive features from modalities such as text, audio, and visual and help in enhancing our model robust performance.
- For MED of utterance level in interactive conversations, we conduct the experiments on both IEMOCAP and AVEC datasets. The results show our model DCWS-RNNs obtains competitive performance in comparison with the state of the art [14] (seen in Table 4 and Table 5).

The remainder of the paper is organized as follows: Section II briefly discusses important previous work; Section III provides the problem definition; Section IV describes our method in detail; Section V provides experimental setup and reports the results; discussion and analysis are presented in Section VI; finally, Section VII concludes the paper.

II. RELATED WORK

Emotion detection [15] involves the intersection of signal processing [16], in-depth learning, edge computing, psychology [17], natural language processing [18], cognitive science [19], and so on [20]. Ekman (1993) [21] launched research on the correlation between emotion and facial cues. Datcu and Rothkrantz (2008, 2011) [22], [23] employed both acoustic information and visual cues for emotion detection. Alm et al. (2005) [24] researched the emotion detection by adding the text-based information, improved in the later work of Strapparava and Mihalcea (2010) [25]. In recent years, MED has become an active research area due to obtaining impetus from researchers. MED means to perform emotion detection from a multimodal learning perspective [3], [26]. A large amount of current research, such as [27]–[31], have been dependent on multimodal techniques.
for emotion detection. These exhibit good performances in emotion detection systems [32]–[34], thus promoting the use of multimodality. Our proposed DCWS-RNNs, in the same way, adopt multimodal (text, audio and visual modalities) fusion approach to represent all the utterances in conversations for performing emotion detection.

Replicating human interaction needs to profoundly understand the conversations. The significant premise is to capture the emotional dynamics. Some important work has pointed out that the emotional dynamics have been looked upon as an interactive phenomenon, rather than being within-person and one-directional [35], [36]. These emotional dynamics can be modeled by observing transition properties, e.g., through investigating the patterns of emotional transition properties [37] or using finite state machines to model transitions [38]. Since the conversation is a temporal event and has a natural temporal nature, contexts of the test utterance play a critical role in emotional detection. Poria et al. [39] utilized contextual information from neighboring utterances of the same speaker to predict emotions. Hazarika et al. [13] calculated the preceding contexts of both parties by employing the memory networks with the same length. Different from these works, we not only use the contexts but divide these contexts into four different aspects.

Memory networks as a storage mechanism, whose memory cells are continuous vectors, are efficient in capturing long-term dependencies [40]. The use of memory networks has been beneficial in multiple research areas, such as speech detection [41], [42], question-answering [43], [44], machine translation [45], commonsense reasoning [46], and others. Zadeh et al. [47] proposed a framework of memory-based sequential learning for multi-view signals. Hazarika et al. [8], [13] both employed memory networks to enable inter-speaker interaction to detect emotions in interactive conversations, with the former combining memories to encode whole utterances and the later using two distinct memory networks. Besides, the attention mechanism and the multiply hop scheme are generally embedded into memory networks. In this paper, these mechanisms are adopted to model the different aspects of contexts of the test utterance.

RNN as a kind of neural network used to process sequence data has also been instrumental in the progress of the MED problem. For example, Poria et al. [39] successfully developed RNN-based deep networks for MED, which was followed by other works [48]–[50]. Majumder et al. [14] proposed an attentive RNN for emotion detection, which is the state of the art in MED. Our work is a follow-up to previous research [13]. Although our proposed model also employs RNN, our method is different for it has four RNNs with different contextual window sizes (corresponding to different time steps) as calculation modules of the contexts.

III. PROBLEM DEFINITION

Our goal is to accurately detect the emotion of utterances present in interactive conversations. Let us define a conversation $U$ to be an asynchronous exchange of utterances between two persons $P_a$ and $P_b$. With $T$ utterances, $U = \{u_1, u_2, \ldots, u_T\}$ denotes a totally ordered set based on temporal occurrence, where $u_i$ is the $i$th utterance at time step $t \in [1, T]$. Here, $u_i$ is spoken by either $P_a$ or $P_b$. The contexts of $u_i$ are those utterances in $U$ except $u_i$. We divide the contexts of $u_i$ into four aspects: own before aspect, the other before aspect, own later aspect and the other later aspect (respectively marked as $A$, $B$, $C$, and $D$). Each aspect is part of the contexts of $u_i$. The corresponding utterances in these four aspects are separately collected and respectively marked as four different sets ($U_A$, $U_B$, $U_C$, and $U_D$). There is no intersection of utterances between every aspect.

Further, considering different contextual window sizes ($W_A$, $W_B$, $W_C$ and $W_D$), these four different sets ($U_A$, $U_B$, $U_C$ and $U_D$) within corresponding window sizes can be respectively constrained to $U_{WA}$, $U_{WB}$, $U_{WC}$ and $U_{WD}$. Concretely, for aspect $A$ of contexts, $U_A$ contains those utterances that occur before $u_i$ and are uttered by the same person who utters $u_i$; $U_{WA}$ contains those utterances that are in $U_A$ and with the farthest distance of $W_A$ from $u_i$. In the same way, $U_B$, $U_{WB}$, $U_C$, $U_{WC}$, $U_D$, and $U_{WD}$ can be respectively obtained referring to aspect $A$.

It should be noted that, at the beginning or the ending of the conversation, the number of the utterances in these constrained sets would have lesser than their corresponding window sizes. So we have the inequality as,

$$|U_{\lambda}^{W_{\lambda}}| \leq W_{\lambda}$$  \hspace{1cm} (1)

where, $\lambda \in \{A, B, C, D\}$.

For brevity, we explain our model using a subscript $\lambda$ which can instantiate to $A$, $B$, $C$ or $D$ in the remaining sections.

| TABLE 1. An example of interactive conversation $U$ with test utterance $u_i = u_{13}$. |
|---|---|
| $U$ | $\{u_1, u_2, u_3, \ldots, u_{23}, u_{25}, u_{26}\}$ |
| $u_i$ | $u_{13}$ |
| $U_A$ | $\{u_1, u_2, u_3, u_7, u_8, u_9, u_{10}, u_{11}\}$ |
| $U_A^{WA}$ | $\{u_4, u_7, u_8, u_{10}, u_{12}\}$ |
| $U_B$ | $\{u_3, u_5, u_6, u_9, u_{11}\}$ |
| $U_B^{WB}$ | $\{u_5, u_6, u_9, u_{11}\}$ |
| $U_C$ | $\{u_{15}, u_{16}, u_{19}, u_{21}, u_{23}, u_{25}, u_{26}\}$ |
| $U_C^{WC}$ | $\{u_{15}, u_{16}, u_{19}\}$ |
| $U_D$ | $\{u_{14}, u_{17}, u_{18}, u_{20}, u_{22}, u_{24}\}$ |
| $U_D^{WD}$ | $\{u_{14}, u_{17}\}$ |

Table 1 provides a sample conversation with different contextual window sizes, $W_A = 9$, $W_B = 8$, $W_C = 6$ and $W_D = 4$. $u_i$ represents the $i$th utterance in $U$. 
IV. DCWS-RNNs MODEL

For improving classification performance, we propose a hypothesis: the contexts of a test utterance $u_t$ should be differentiated. Further, we utilize different contextual window sizes to represent the implicit weights of different aspects of contexts. Based on this assumption, we develop a model named DCWS-RNNs. The overall working process of our model is as follows.

Firstly, four different sets ($U^A$, $U^B$, $U^C$, and $U^D$) relevant to the test utterance $u_t$ are separately collected and respectively sent to four RNNs to generate four initial memories, with each RNN accepting a corresponding set. These RNNs have different contextual window sizes. One RNN is assigned to one aspect and each RNN works independently. The RNN can model its corresponding aspect into memory cells using GRUs. Then, these memories are conveyed to the attention block to perform an attention mechanism. After attention operation, all the memories of four aspects are accumulated and merged with $u_t$. This produces an updated representation of the test utterance. Furthermore, a multiple hop scheme is adopted to conduct the iterations. At each hop, the output of the RNN module is used as its input in the next hop. The ultimate emotion representation of $u_t$ is obtained after $H$ hops, which is used for the emotion classification. Figure 2 illustrates the overall model.

In the following subsections, we will detail the DCWS-RNNs model.

A. SEPARATE GRUs MODULES

For the computations of different aspects of contexts, we employ four independent RNNs that are based on GRUs. Similar to an LSTM [51] in RNN, the GRU is a simpler gating mechanism with similar computation performance and introduced by Cho et al. [52]. At any time step $t$ of a temporal sequence, the GRU controls the combination criteria with the $t$th input $u_t$ and previous state $s_{t-1}$ by computing two gates, $r_t$ (reset gate) and $z_t$ (update gate). The computations are:

$$z_t = \sigma (V^z \cdot u_t + W^z \cdot s_{t-1} + b^z)$$  \hspace{1cm} (2)
$$r_t = \sigma (V^r \cdot u_t + W^r \cdot s_{t-1} + b^r)$$  \hspace{1cm} (3)
$$h_t = \tanh(V^h \cdot u_t + W^h \cdot (s_{t-1} \odot r_t) + b^h)$$  \hspace{1cm} (4)
$$s_t = (1 - z_t) \odot h_t + z_t \odot s_{t-1}$$  \hspace{1cm} (5)

Here, $V$, $W$ and $b$ are trainable parameters and $\odot$ represents element-wise multiplication. Specially, we adopt the rules: for before utterances of $u_t$, the direction is starting from the oldest one and; for later utterances of $u_t$, the direction is starting from the farthest one.

The initial memories of four different aspects of contexts are computed using four RNNs which are respectively based on separate GRU$^A$,$ GRU^B$, GRU$^{init}$ and GRU$^D$ (Init means initialization).

For the memory representation of $A$ aspect, marked as $M^A$, the utterances in set $U^A$ are framed as a sequence based on temporal occurrence (starting from the oldest one) and fed to the GRU$^{init}$-based RNN to produce $M^A$. At each time step $i \in [t - W_A, t - 1]$, the internal state $s_i$ forms the $i$th memory cell of $M^A$.

For the memory representation of $B$ aspect, marked as $M^B$, the utterances in set $U^B$ are framed as a sequence based on temporal occurrence (starting from the oldest one) and fed to the GRU$^{init}$-based RNN to produce $M^B$. At each time step $i \in [t - W_B, t - 1]$, the internal state $s_i$ forms the $i$th memory cell of $M^B$.

For the memory representation of $C$ aspect, marked as $M^C$, the utterances in set $U^C$ are framed as a sequence based on temporal occurrence (starting from the farthest one) and fed to the GRU$^{init}$-based RNN to produce $M^C$. At each time step $i \in [t + 1, t + W_C]$, the internal state $s_i$ forms the $i$th memory cell of $M^C$.

For the memory representation of $D$ aspect, marked as $M^D$, the utterances in set $U^D$ are framed as a sequence based on temporal occurrence (starting from the farthest one) and fed to the GRU$^{init}$-based RNN to produce $M^D$. At each time step $i \in [t + 1, t + W_D]$, the internal state $s_i$ forms the $i$th memory cell of $M^D$.

B. INDEPENDENT ATTENTION MECHANISM

In this module, the test utterance $u_t$ is first embedded into a vector $q(1)$ of dimension $R^u_t$ using a projection matrix $P$. Next, $q(1)$ and $M^A$ are read by independent attention module. This module performs an independent attention mechanism on them, resulting in an attention vector $p(1)$ corresponding to $\lambda$. The dimension of $p(1)$ is $R_W$. The computations are:

$$q(1) = P \cdot u_t$$  \hspace{1cm} (6)
$$p(1) = \text{softmax}((M^A)^T \cdot q(1))$$  \hspace{1cm} (7)

where, $\text{softmax}(x_j) = e^{x_j} / \sum_k e^{x_k}$, and the attention vector $p(1)$ is a probability distribution over $M^A$. The $j$th normalized score of $p(1)$ represents the relevance of $j$th memory cell with respect to the test utterance. The $p(1)$ and $M^A$ are then used to find a weighted representation. We do this by taking an inner product as follows:

$$\text{att}M^A = M^A \cdot p(1)$$  \hspace{1cm} (8)

This representation $\text{att}M^A$ contains a weighted contextual summary accumulated from the memories of the aspect that is corresponding to $\lambda$. As $\lambda$ can instantiate to $A$, $B$, $C$ or $D$, the representation of the test utterance can be updated by adding all weighted representations as:

$$q(2) = q(1) + \text{att}M^A + \text{att}M^B + \text{att}M^C + \text{att}M^D$$  \hspace{1cm} (9)

C. MULTIPLE HOPS

For better performance, we perform $H$ hops on all initial memories (including $M^A$, $M^B$, $M^C$ and $M^D$). Adopting a multiple hop scheme is inspired by several recent works [40], [44], which suggest the importance of multiple input/output iterations for performing transitive inference. The multiple hop scheme over the inputs is crucial to...
achieving high performance in our experiments. The iterative nature of this scheme can enable the model to focus on relevant parts of the inputs. It also allows for a type of transitive inference. Each iteration provides our model with newly relevant information about the inputs. In other words, some information that may be ignored previous iterations...
can be retrieved during the iterations. Multiple hops also are instrumental in improving the focus of attention heads which might miss essential contexts in a single hop. This scheme is done as follows:

- Each memory representation $M^{(h)}_{\lambda}$ ($h$ is an integer greater than zero) is updated by using dynamic RNN whose cell is $GRU_\lambda$. The output at the $h^{th}$ hop is used as the input of next-hop ($h + 1)^{th}$. This constraint of sharing parameters adjacently between hops is added for the reduction in total parameters and ease of training. The update operation is as:

$$M^{(h+1)}_{\lambda} = GRU_\lambda(M^{(h)}_{\lambda})$$ (10)

- At every hop, the representation of the test utterance is updated by the equation as:

$$q^{(h+1)} = q^{(h)} + attM^{(h)}_A + attM^{(h)}_B + attM^{(h)}_C + attM^{(h)}_D$$ (11)

- After $H$ hops, the representation $q^{(H+1)}$ is sent to the final prediction.

D. FINAL PREDICTION

Through an affine transformation, the final representation vector can be obtained using the equation as:

$$\hat{y} = softmax(W_o \cdot (q^{(H+1)}) + b_o)$$ (12)

For classification training, categorical cross-entropy loss along with L2 regularization is used as the cost measure:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} \log \hat{y}_{i,j} + \alpha \| \theta \|_2$$ (13)

Here, $N$ denotes total utterances across all videos and $C$ is the number of emotion categories. $y_{i,j}$ is the ground truth value of $i^{th}$ utterance of belonging to class $j$ from the training set. $\hat{y}_{i,j}$ is its predicted probability of emotion labels of belonging to class $j$. $\alpha$ is the L2-regularizer weight, and $\theta$ is the set of trainable parameters.

For regression, $\hat{y}$ is a scalar (without softmax normalization) whose scores are used to calculate the mean squared error (MSE) cost metric.

Figure 3 presents a flowchart of DCWS-RNNs.

V. EXPERIMENTS

A. DATASET DETAILS

We perform experiments on two emotion detection datasets: the interactive emotional dyadic motion capture database (IEMOCAP)\(^1\) [53] and the 2012 Audio-Visual Emotion Challenge and Workshop dataset (AVEC 2012) [54]. IEMOCAP provides rich video and audio samples for all the utterances along with transcriptions and, generally, is regarded as a benchmark of emotion detection. It contains 12 hours of dyadic conversational videos. These videos are grouped into five sessions and split into 5 minutes of interaction between pairs of 10 unique speakers (5 male and 5 female). Each pair is assigned to diverse multiple conversation scenarios of dialogues. All the conversations are segmented into spoken utterances. Each utterance is tagged by at least 3 annotators with one of the labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and others. In our experiments, however, we consider the first six categories with majority agreement (i.e., at least two out of three annotators labeled the same emotion), to compare with the state-of-the-art [14] and other works. Table 2 presents the class distribution in the IEMOCAP dataset.

| Class | ang. | hap. | sad. | neu. | exc. | fru. |
|-------|------|------|------|------|------|------|
| Number | 1103 | 648  | 1084 | 1708 | 1041 | 1849 |

\(^1\)https://sail.usc.edu/iemocap/

AVEC [54] is a modification of SEMAINE database [55] containing human-agent interactions, which requires...
predictions of four real valued affective attributes: valence \((-1, 1]\), arousal \((-1, 1]\), expectancy \((-1, 1]\), and power \([0, \infty]\). The annotations are available for every 0.2 seconds in each video. However, in order to adapt our problem definition, the attributes over the span of an utterance are averaged as the utterance-level annotation.

**B. MULTIMODAL FEATURE DATA**

For a fair comparison with the state-of-the-art method [14], we employ the identical feature data that is downloaded from their public website.\(^2\) The extraction procedures of feature data are briefly described below.

The textual features are extracted from the transcript of each utterance through a simple CNN with a single convolutional layer followed by max-pooling [56]. The convolutional layer consists of three distinct convolution filters, followed by max-pooling and rectified linear unit (ReLU) activation. These activations are concatenated and then fed to a fully connected layer to obtain the textual utterance representation \(t_u\).

The audio features are obtained by using the open-sourced software open SMILE. Specifically, the IS13_ComParE\(^3\) extractor can provide 6373 features for each utterance. By using a fully-connected layer, the dimension is reduced and the audio feature vector \(a_u\) is provided.

The visual features are captured by using 3D-CNN [57]. Outputs of this 3D convolution are then subjected to max-pooling followed by a fully connected layer. Finally, the activations of this layer form the video representation \(v_u\).

By concatenating the above three multimodal features, the final representation of an utterance is \(u = [t_u, a_u, v_u]\). This multimodal representation is adopted for all utterances in a conversation.

For the details of extraction procedures, readers can refer to the literature [8], [13], [14].

**C. TRAINING DETAILS**

A large section of research works in MED is performed on datasets where training and test splits may share certain speakers. However, as each individual has a unique way of expressing emotions in the real-world, it is crucial to develop a model that can find generic and person-independent features. Hence, we perform person independent experiments to test our method. We partition the dataset into training and test sets with roughly 80/20 ratio such that the partitions do not share any speaker (i.e. on the IEMOCAP dataset, only the first eight speakers from session 1 to 4 belong to the training set, while the last two speakers from session 5 belong to the test set). Furthermore, 20% of the training set is used as a validation set for hyper-parameters tuning. Table 3 summaries the splits of both datasets.

For optimization of the experimental parameters on the IEMOCAP dataset, we use the Adam optimizer [58], which starts with an initial learning rate (LR) of 0.01, to train our network. Termination of the training phase is decided by early-stopping with the patience of 10 epochs through monitoring the validation loss. The network is subjected to regularization in the form of Dropout [59] and Gradient-clipping for a norm of 40. The final hyper-parameters are decided using a random search [60]. Based on validation performance, contextual window sizes of \(W_A\), \(W_B\), \(W_C\) and \(W_D\) are respectively set to be 31, 22, 24 and 17. The number of hops \(H\) is fixed at 3. The embedding dimension size \(R^d\) is set as 50.

**D. BASELINES**

- **TFN** [61] captures inter-modality and intra-modality dynamic interactions by using the tensor outer product. In their model, inter-modality dynamics are modeled with a new multimodal fusion approach, named Tensor Fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. The intra-modality dynamics are modeled through three Modality Embedding Sub-networks, for language, visual and acoustic modalities, respectively.

- **CATF-LSTM** [62] proposes attention-based networks for improving both context learning and dynamic feature fusion. It utilizes an attention-based fusion mechanism to amplify the higher quality and informative modalities during fusion in multimodal classification, followed by a contextual attention-based LSTM network that models the contextual relationship among utterances and prioritizes the important contextual information for classification.

- **CMN** [13] models separate contexts from conversational histories using two distinct GRUs for two speakers. These contexts are stored as memories and an attention mechanism is performed on them. The weighted memories are then merged with the test utterance using an addition operation. Furthermore, this merged representation is used to classify its emotion category after the multiple hop scheme that is done by the repeated input and output cycle.

- **ICON** [8] considers the preceding utterances of both speakers falling within a context-window and models their self-emotional influences using local GRUs. Furthermore, to incorporate inter-speaker influences, a global representation is generated using a global GRU that is considered as a memory to track the overall conversational flow.

- **DialogueRNN** [14] is a state-of-the-art method for MED. The method is based on the assumption that

\(^2\)https://github.com/senticnet
\(^3\)http://audeering.com/technology/opensmile
there are three major aspects relevant to the emotion in a conversation: the speaker, the context from the preceding utterances, and the emotion of the preceding utterances. It employs a recurrent network based on three GRUs to track these aspects. The global GRU and party GRU are respectively used to update the global context and the individual speaker states during the conversation. Further, the emotion GRU is employed to track emotional state through the conversation.

However, the contexts are not considered in the method TFN; the speakers in the conversations are not differentiated in the method CATF-LSTM; in the models, CMN, ICON, and DialogueRNN, the different aspects of contexts are not distinguished and their weights are not represented.

For the performance comparison, we reimplement the above methods in our experiments.

### TABLE 4. Performance of DCWS-RNNs in terms of Accuracy and F1-score on the IEMOCAP dataset.

| Models        | Accuracy | F1   |
|---------------|----------|------|
| CATF-LSTM     | 58.04    | 57.54|
| TFN           | 58.23    | 57.93|
| CMN           | 60.94    | 60.78|
| ICON          | 61.18    | 61.21|
| DialogueRNN   | 62.05    | 61.62|
| DCWS-RNNs     | 64.45    | 62.43|

### E. RESULTS

On the IEMOCAP dataset, we mainly conduct the static analysis of the performance of our proposed model DCWS-RNNs. Thus, for detecting the emotion of a test utterance $u_t$, we can use those utterances that occur before or later $u_t$ to improve the emotion representation. We use Accuracy and F1-score to evaluate classification performance. We experiment on the IEMOCAP dataset and compare DCWS-RNNs with the above baseline methods. The results are summarized in Table 4. As shown in Table 4, our model DCWS-RNNs outperforms the compared models with classification performance increase in Accuracy ranging from 2.40% to 6.41% and in F1 score ranging from 0.81% to 4.89%. Especially, our method DCWS-RNNs succeeds over the state of the art DialogueRNN by 2.40% in Accuracy. This suggests that the proposed DCWS-RNNs model is more capable of capturing contextual information.

Furthermore, for a probability to apply in the real-time dialogue system, we perform the experiments on the AVEC dataset just only using the history information (both A aspect of contexts and B aspect of contexts). We mark this variant model as DCWS-RNNs*. The results on the AVEC dataset are presented in Table 5. Our DCWS-RNNs* obtains competitive performance in comparison with the state of the art DialogueRNN.

### VI. DISCUSSION

#### A. HYPER-PARAMETERS

In consideration of the combination explosion, we use the random search to set the values of $W_A, W_B, W_C$ and $W_D$. In the experiments, we find that different contextual window sizes have a great impact on the performance of the model. This verifies our hypothesis that the contexts of a test utterance should be divided into different aspects and, the weights of different aspects of contexts can be represented with different contextual window sizes. The improvement of setting these values remains as our future work. On the other hand, we plot the performance trends of model DCWS-RNNs on the IEMOCAP dataset concerning the main hyper-parameter $H$ when fixing the values of $W_A = 31$, $W_B = 22$, $W_C = 24$ and $W_D = 17$. As demonstrated in Figure 4, the performance initially improves, which is showing the importance of the multiple hop scheme. However, with a further increase, the performance decrease. Finally, the best performance is obtained at $H = 3$ in our experiments.

#### B. MULTIMODALITY

Here we give a short multimodal analysis. Table 6 summarizes the performance for different combinations of modes used by DCWS-RNNs on IEMOCAP and DCWS-RNNs* on AVEC. As seen in the table, we observe the trend in
our experiments where the trimodal network outperforms a unimodal network. Besides, text modality performs best out of the three, which is reaffirming its significance in multimodal systems. Although other modalities contribute to improving the performance, that contribution is little compared to the textual modality, which is aligned with our expectations.

**C. Case Studies**

In order to understand how DCWS-RNNs pays attention to different aspects of contexts of a test utterance after 3-hops, we manually explore utterances in the test set of IEMOCAP and perform visualization of the $p_\lambda$ attention vector (Eq. (7)) over them. Figure 5 presents four cases which are corresponding to four different aspects of contexts and, whose utterances are respectively assigned to the highest attention score. Each case has its own range of attention scores, from the lowest 0 value to the highest value. The symbol $u_t$ denotes the test utterance and $u_d$ denotes the dependent utterance that is most relevant to the test utterance and assigned to the highest attention score.

As shown in Figure 5 (a), it displays the important role of self-emotional influences and how DCWS-RNNs pays the highest attention to the utterance that is in $A$ aspect of contexts. In this case, the model DCWS-RNNs anticipates the test utterance (turn 6) correctly by attending to the utterance (turn 3) that is in $A$ aspect of contexts and assigned to the highest attention score. This result indicates the presence of long-term emotional dependencies and the need to consider the $A$ aspect of contexts. In Figure 5 (b), it visualizes the important role of inter-speaker influences and how DCWS-RNNs processes dependency on the utterances that are in $B$ aspect of contexts. In this case, the model DCWS-RNNs predicts the test utterance (turn 2) correctly by attending to the utterance (turn 1) that is in $B$ aspect of contexts and assigned to the highest attention score. This result indicates the need to consider the $B$ aspect of contexts.

Figure 5 (c) demonstrates the ability of DCWS-RNNs to model self-emotional influences and how it pays the highest attention to the utterance that is in $C$ aspect of contexts. In this case, while classifying the test utterance (turn 63), DCWS-RNNs finds contextual information from $C$ aspect of contexts and focuses more on the utterance (turn 65). This utterance is most relevant to the test utterance and assigned to the highest attention score. This result indicates the benefit of considering the $C$ aspect of contexts to capture emotional

**FIGURE 5.** Cases analysis of highest attention score over different attention vectors.
correlations across time. Figure 5 (d) shows the important role of inter-speaker influences and how DCWS-RNNs differentiates the utterances that are in D aspect of contexts. In this case, while classifying the test utterance (turn 51), DCWS-RNNs finds contextual information from D aspect of contexts and focuses more on the utterance (turn 65) that is most relevant to the test utterance and assigned to the highest attention score. This result indicates the benefit of considering the D aspect of contexts to capture emotional correlations across time.

From the above cases, we can obtain evidence that the weights of different aspects of contexts can be represented with different contextual window sizes. In other words, all the different aspects of contexts unequally contribute to the output representation. This helps to capture the more relevant information of the test utterance from the perspective of emotional contexts, making our model performs better than the state of the art DialogueRNN.

**TABLE 7. Ablation study for modules of DCWS-RNNs on IEMOCAP.**

| DCWS-RNNs variants | IEMOCAP |
|---------------------|---------|
| aspect A | aspect B | aspect C | aspect D | Accuracy | F1 |
| - | + | + | + | 57.92 | 56.99 |
| + | - | + | + | 60.73 | 58.13 |
| + | + | - | + | 61.98 | 61.86 |
| + | + | + | - | 62.42 | 60.67 |

**D. ABLATION STUDY**

The main novelty of our DCWS-RNNs model is to divide the contexts of $u_t$ into four different aspects: A, B, C, and D. These aspects have been described in section III. To check the impact of the different aspects of contexts present in DCWS-RNNs, we conduct an ablation study by removing one of the aspects at a time and evaluate the performance on the IEMOCAP dataset. Table 7 demonstrates the results of this study. As expected, A aspect of contexts stands most important, as without its presence the performance falls by 6.53%. We believe that A aspect of contexts contains more correlation information that is relevant to the test utterance in comparison with B, C, and D aspect. The B aspect of contexts is also important, but less than A aspect as its absence causes performance to fall by 3.70%. We speculate the reason to be the self-emotional influence is more important than inter-speaker emotional influence.

**VII. CONCLUSION**

In this paper, we have presented DCWS-RNNs, a multimodal model for emotion detection in interactive conversations. In contrast to the state-of-the-art method, DialogueRNN, our method considers the contexts of a test utterance should be divided into four aspects. Further, the implicit weights of these different aspects of contexts can be represented with different contextual window sizes, which can help to improve accuracy performance. Our experimental results confirm the performance of DCWS-RNNs with such a modeling scheme. In the future, we plan to investigate the time complexity of our model and newer structures of RNNs.

**REFERENCES**

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.
[2] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Netw., vol. 61, pp. 85–117, Jan. 2015.
[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” Inf. Fusion, vol. 37, pp. 98–125, Sep. 2017.
[4] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” Knowl.-Based Syst., vol. 161, pp. 124–133, Dec. 2018.
[5] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” IEEE Access, vol. 7, pp. 10,0943–10,0953, 2019.
[6] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” 2018, arXiv:1810.02508. [Online]. Available: http://arxiv.org/abs/1810.02508
[7] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” in Proc. 32nd AAAI Conf. Intell., 2018, pp. 1–9.
[8] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, “ICON: Interactive conversational memory network for multimodal emotion detection,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 2594–2604.
[9] M. W. Morris and D. Keltner, “How emotions work: The social functions of emotional expression in negotiations,” Res. Organizational Behav., vol. 22, pp. 1–50, 2000.
[10] F. Liu and S. Maitlis, “Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings,” J. Manage. Stud., vol. 51, no. 2, pp. 202–234, Mar. 2014.
[11] P. Koval and P. Kuppens, “Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia,” Emotion, vol. 12, no. 2, p. 256, 2012.
[12] P. Kuppens, N. B. Allen, and L. B. Sheeber, “Emotional inertia and psychological maladjustment,” Psychol. Sci., vol. 21, no. 7, pp. 984–991, Jul. 2010.
[13] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, “Conversational memory network for emotion recognition in dyadic dialogue videos,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., vol. 1, 2018, pp. 2212–232.
[14] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “DialogueRNN: An attentive rnn for emotion detection in conversations,” in Proc. AAAI Conf. Intell., vol. 33, 2019, pp. 6818–6825.
[15] M. Abdul-Mageed and L. Ungar, “EmoNet: Fine-grained emotion detection with gated recurrent neural networks,” in Proc. 55th Ann. Meeting Assoc. Comput. Linguistics, vol. 1, 2017, pp. 718–728.
[16] S. S. Bhattacharyya, E. F. Deprettere, R. Leupers, and J. Takala, Handbook of Signal Processing Systems. Cham, Switzerland: Springer, 2018.
[17] H. Coolican, Research Methods and Statistics in Psychology. London, U.K.: Psychology Press, 2017.
[18] Y. Goldberg, “Neural network methods for natural language processing,” Synth. Lectures Hum. Lang. Technol., vol. 10, no. 1, pp. 1–309, Apr. 2017.
[19] A. Collins and D. G. Bobrow, Representation and Understanding: Studies in Cognitive Science. Amsterdam, The Netherlands: Elsevier, 2017.
[20] M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, “Audio-visual emotion recognition using big data towards SG,” Mobile Netw. Appl., vol. 21, no. 5, pp. 753–763, Oct. 2016.
[21] P. Ekman, “Facial expression and emotion,” Amer. Psychol., vol. 48, no. 4, p. 384, 1993.
[22] D. Datcu and L. J. M. Rothkrantz, “Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition,” Hoboken, NJ, USA: Wiley, 2008.
[23] D. Datcu and L. J. M. Rothkrantz, “Emotion recognition using bimodal data fusion,” in Proc. 12th Int. Conf. Comput. Syst. Technol. (CompSyst-Tech), 2011, pp. 122–128.
[24] C. O. Alin, D. Roth, and R. Sproat, “Emotions from text: Machine learning for text-based emotion prediction,” in Proc. Conf. Hum. Lang. Technol. Empirical Methods Natural Lang. Process., 2005, pp. 579–586.
[25] C. Strapparava and R. Mihalcea, “Annotating and identifying emotions in text,” in Intelligent Information Access. Berlin, Germany: Springer, 2010, pp. 21–38.
[26] F. Alonso-Martín, M. Malfaz, J. S. Queeira, J. F. Gorostiza, and M. A. Salicrú, “A multimodal emotion detection system during human–robot interaction,” Sensors, vol. 13, no. 11, pp. 15549–15581, 2013.
W. Lim, D. Jang, and T. Lee, “Self-attentive feature-level fusion for multimodal emotion detection,” in Proc. IEEE Conf. Multimedia Inf. Process. Ret. (MIPR), Apr. 2018, pp. 196–201.

C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegzryn-Wolska, “Survey on AI-based multimodal methods for emotion detection,” in High-Performance Modelling and Simulation for Big Data Applications, Cham, Switzerland: Springer, 2019, pp. 307–324.

A. Kashii, K. Takashio, and H. Tokuda, “Ex-amp robot: Expressive robotic avatar with multimodal emotion detection to enhance communication of users with motor disabilities,” in Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN), Aug. 2017, pp. 864–870.

D. Abin, S. Saini, R. Rao, V. Vaichole, and A. Rane, “Survey of various approaches of emotion detection via multimodal approach,” Int. Rev. J. Eng. Technol., vol. 4, no. 12, p. 287, 2017.

H. Ranganathan, S. Chakraborty, and S. Panchanathan, “Multimodal emotion recognition using deep learning architectures,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2016, pp. 1–9.

M. S. Hossain and G. Muhammad, “An emotion recognition system for mobile applications,” IEEE Access, vol. 5, pp. 2281–2287, 2017.

S. Turabzadeh, H. Meng, R. Swash, M. Pleva, and J. Juhar, “Facial expression emotion detection for real-time embedded systems,” Technologies, vol. 6, no. 1, p. 17, Jan. 2018.

J. Herzig, D. Konopnicki, T. Sandbank, and M. Shmueli-Scheuer, “Emotion detection and expression integration in dialog systems,” U.S. Patent 10 372 825, Aug. 6, 2019.

J. M. Richards, E. A. Butler, and J. J. Gross, “Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings,” J. Social Pers. Relationships, vol. 20, no. 5, pp. 599–620, Oct. 2003.

S. Hareli and A. Rafaeli, “Emotion cycles: On the social influence of emotion in organizations,” Res. Organizational Behav., vol. 28, pp. 35–59, Jan. 2008.

L. Yang, F. L. Hong, and W. GUO, “Textbased emotion transformation analysis,” in Proc. Acess, vol. 9, p. 026, Sep. 2011.

P. Xiaolan, X. Lun, L. Xin, and W. Zhiliang, “Emotional state transition model based on stimulus and personality characteristics,” China Commun., vol. 10, no. 6, pp. 146–155, Jun. 2013.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2017, pp. 173–883.

J. Weston, S. Chopra, and A. Bordes, “Memory networks,” 2014, arXiv:1410.3916. [Online]. Available: http://arxiv.org/abs/1410.3916

H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,” IEEE Access, vol. 6, pp. 13825–13835, 2018.

O. Faust, A. Shenfield, M. Kareem, T. R. San, H. Fujita, and U. R. Acharya, “Automated detection of atrial fibrillation using long short-term memory network with RR interval signals,” Comput. Biol. Med., vol. 102, pp. 327–335, Nov. 2018.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu, “Multi-scale attentive interaction networks for chinese medical question answer selection,” IEEE Access, vol. 6, pp. 74061–74071, 2018.

A. Kumar, O. Irsy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhigang, R. Paulus, and R. Socher, “Ask us anything: Dynamic memory networks for natural language processing,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 1378–1387.

Y. Heo, S. Kang, and D. Yoo, “Multimodal neural machine translation with weakly labeled images,” IEEE Access, vol. 7, pp. 54042–54053, 2019.

S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, “EmoSenticSpace: A novel framework for affective common-sense reasoning,” Knowl.-Based Sys., vol. 69, pp. 108–123, Oct. 2014.

A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “Memory fusion network for multi-view sequential learning,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–8.

W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Dec. 2016, pp. 1–4.

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

Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” in Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI), 2016, pp. 445–450.

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1800, 1997.

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” 2014, arXiv:1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Lang. Resour. Eval., vol. 42, no. 4, pp. 335, Dec. 2008.

B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, “AVEC 2012: The continuous audio/visual emotion challenge,” in Proc. 14th ACM Int. Conf. Multimodal Interact., 2012, pp. 449–456.

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 5–17, Jan. 2012.

Y. Kim, “Convolutional neural networks for sentence classification,” 2014, arXiv:1408.5882. [Online]. Available: http://arxiv.org/abs/1408.5882

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” J. Mach. Learn. Res., vol. 13, pp. 281–305, Feb. 2012.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” 2017, arXiv:1707.07250. [Online]. Available: http://arxiv.org/abs/1707.07250

S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, “Multi-level multiple attentions for contextual multimodal sentiment analysis,” in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2017, pp. 1033–1038.

HELHAI LAI received the M.S. degree from South China Normal University, Guangzhou, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His current research interests include sentiment analysis, neural networks, and service computing theory.

HELINGYING CHEN received the M.S. and Ph.D. degrees from the Guangdong University of Technology, Guangzhou, China, in 2002 and 2007, respectively. She is currently an Associate Professor with South China Normal University. Her current research interests include artificial intelligence, machine learning, and data mining.

SHUANGYAN WU received the M.S. degree from South China Normal University, Guangzhou, China, in 2017. Her current research interests include deep learning, image processing, and data mining.