INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speaker-aware Cross-modal Fusion Architecture for Conversational Emotion Recognition

Huan Zhao[1], *Bo Li*[1,*], *Zixing Zhang*[1]

[1]College of Computer Science and Electronic Engineering, Hunan University, China

{hzhao, blee, zixingzhang} @hnu.edu.cn

Abstract

Conversational Emotion Recognition (CER) is an important topic in the construction of intelligent human-machine interaction systems. The emotion is mainly influenced by the conversational context and the speakers. In addition, sufficient utilization of the relevant features of both speech and text modes is also crucial to the performance of CER. Based on the above considerations, we propose a novel Speaker-aware Cross-modal Fusion Architecture (SCFA). Within a single modality, we design a conversation encoder, including a context encoder and a speaker-aware encoder, to model the conversational content and the intra- and inter-speaker influence, respectively. On this basis, cross-modal fusion attention is introduced to extract the cross-modal characteristics of the conversation, so as to better detect the emotions in conversation. We conduct experiments on the IEMOCAP and MELD datasets. Compared with state-of-the-art baselines, SCFA achieves better performance on average.

Index Terms: Speech emotion recognition, Multi modal feature fusion, Attention mechanism

1. Introduction

Emotion awareness is an important property of advanced artificial intelligence. Conversational emotion recognition(CER) which aims to identify the emotional status of utterances in conversations has been one of the research hotspots[1, 2, 3, 4]. It makes human-machine interaction more natural. Conversations contain information in both speech and text models. Since speech is the acoustic expression of semantics, researchers respectively analyze the emotional information of speech from audio and text, using speech processing and natural language processing[5, 6, 7]. However, previous work has proved that it cannot accurately detect the emotional category through one modal information[8, 9, 10, 11, 12, 13]. For example, from the text aspect, it is hard to distinguish if the sentence "I just deleted my Twitter" is an expression of happiness or anger. In this case, tone can be a good aid in judging emotions. Therefore, models must be capable of processing cross-modal information. Because of the ability to capture the information of different modalities, cross-modal models can achieve better performance than uni-modal models.

Since conversation is an interactive process, an important feature of CER that differs from traditional emotion recognition tasks is the influence of conversation participants. On the one hand, the speaker's emotion has continuity[14], that is, his own conversational context will affect the expression of the current discourse emotion. On the other hand, the emotions of other speakers will have an impact on the current speaker, thus making his mood change[15]. Many studies [16, 17, 18, 19] have shown that modeling interactions between speakers can better extract conversational information An effective approach is to construct relationship graphs to model the dependencies between speakers, and then use graph neural networks to process them[19]. However, such models are difficult to handle when the number of speakers changes, and the increasing number of speakers will make the model very complicated. In order to solve these problems, [20] proposes a method to simplify speaker dependence. Instead of setting relationships between every two speakers, they establish dependency relationships with all other speakers centered on the current target speaker. It should be noted that excessive processing of conversational context based on speaker dependence may lead to the loss of some semantic information.

In order to solve the above problem and improve the performance of CER, we propose a novel Speaker-aware Cross-modal Fusion Architecture (SCFA), which consists of three modules: utterance encoding, intra-modal conversation encoding, and cross-modal fusion. First, at the discourse level, the input of the two modes is embedded, and the utterance-level features are obtained by an utterance-level encoder, respectively. Those utterance-level features are concatenated to obtain conversational context. Next, within a single mode, a context transformer is used to combine the conversational context with the concatenated discourse original embeddings to maximize the retention of conversational semantic information. Meanwhile, a speaker-aware transformer is employed to model the intra- and inter-speaker dependencies. The intra-modal conversational encoding is obtained by combining the above two features using intra- modal attention. Then, cross-modal fusion attention is introduced to make full use of the information interaction between the modes to obtain the cross-modal session features. Finally, intra-modal and cross-modal features are fed into a classifier to predict the emotion category of utterances.

The contributions of this paper are summarized as follows

- For intra- modal conversational feature, we propose a conversational encoder, which consists of a context encoder and a speaker-aware encoder. The former is used to model the context feature of the conversation, while the latter is used to capture the intra- and inter-speaker influence of the emotion.
- We introduce a Cross-modal Fusion Attention (CFA) to model the interaction of conversational features between modes.
- Experimental results on IEMOCAP and MELD datasets show that the average results of the proposed method achieve better performance on the two datasets.

2. Methodology

We propose a speaker-aware cross-modal fusion approach for conversational emotion recognition. The framework of the

10.21437/Interspeech.2023-1650

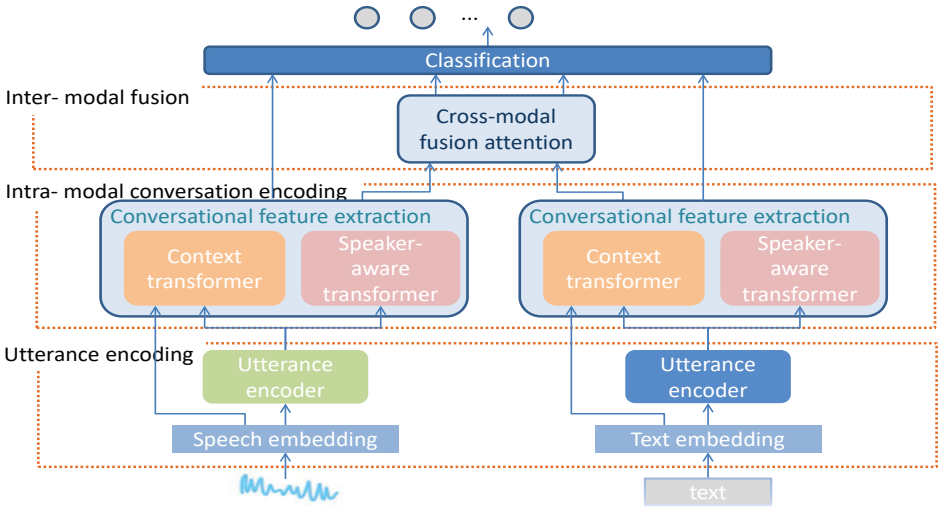

Figure 1: *The framework of the proposed speaker-aware cross-modal fusion architecture(SCFA). Given two input modes of text and speech, the utterance encoding module firstly encodes the input of the two models and extracts the discourse features respectively. The extracted discourse features are spliced into the conversational context and then input into the intra-modal conversation encoding module. We introduce a context transformer and a speaker-aware transformer respectively. The former is used to capture the global features of the conversational context, and the latter captures the speakers themselves and the interactions between speakers by means of masks. The above features are then entered into a CFA to extract the cross-modal fusion features. Finally, an emotion classifier is used to identify the emotional labels of each utterance.*

model is shown in Figure 1. In the encoder module, we encode the input data of two modes to obtain the feature representation in the word level. In the speech coding module, the speech and text are embedded and appropriate networks are used to extract the utterance features in the two modes, respectively. The utterance-level features are concatenated to obtain the conversational context features, which will be sent to the next module for further processing. In the intra-modal conversation encoding module, we design a feature extraction network that combines speaker-aware and conversational context information. Among them, context transformers combine the original embedding and global characteristics of a conversation to fully capture the context of a conversation. speaker-aware transformers employ mask mechanisms to model the intra- and inter-speaker interactions. Through the above modules, we obtain the conversation-level features in speech and text modes. Then, we introduce cross-modal fusion attention to capture cross-modal conversation features. Finally, the intra- and cross-modal conversational features are fed into the classifier to predict the emotional category.

2.1. Task Definition

Formally, a spoken conversation can be defined as $C = \{u_1, u_2, \cdots, u_N\}$, where N is the number of utterances in the conversation C. Utterance $u_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,L_i}\}$, where$L_i$ is the number of words in u_i. The conversation involves M speakers denoted as $SP = sp_1, sp_2, \cdots, sp_M$, where M is the number of speakers. The task of conversational emotion recognition is to predict the emotion label $Y = \{y_1, y_2, \cdots, y_N\}$ for each utterance correspondingly.

2.2. Utterance encoding

Text encoding Before feature extraction, we use word embedding to represent the input utterances and connect them to get the original embedding representation of the conversation.

$$e_{i,j} = embedding(w_{i,j}) \tag{1}$$

$$u_i^o = \{e_{i,1}, \cdots, e_{i,L_i}\}, \tag{2}$$

where $e_{i,j} \in \mathbb{R}^{1*d_e}$ is the embedding of $w_{i,j}$, d_e is the dimension of $e_{i,j}$. In the same way, $u_i^o \in \mathbb{R}^{L_i*d_e}$ and is the embedding of u_i. After that, BiGRUs are applied to encode the text features. Each GRU unit calculates a hidden state defined as

$$\begin{aligned} \overrightarrow{h^t}, \overleftarrow{h^t} &= BiGRU(\overrightarrow{h^{t-1}}, \overleftarrow{h^{t+1}}, x^t) \\ h^t &= concat(\overrightarrow{h^t}, \overleftarrow{h^t}), \end{aligned} \tag{3}$$

where x^t is the input of the current moment t, h^{t-1} (h^{t+1}) is the hidden state of the BiGRU at the previous moment $t-1$ ($t+1$). After utterance feature extraction and concatenation, the conversational context can be represented as $H^t = \{h_1^t, h_2^t, \cdots, h_N^t\}$.

Speech embedding We employ 3D static representation [21] to extract the feature of the speech signal. After a series of operations, containing Fourier transform, Mel spectrum filtering, and derivation, the 3D static features of the speech signal composed of log-Mel, δ, and $\delta - \delta$ can be obtained. In order to capture utterance-level audio features, we adopt local feature learning blocks (LFLB)[22], which are composed of multiple 2-dimensional convolutional layers, batch normalization layers, and Leaky ReLu layers. Thus, the corresponding word-level audio feature sequence $h_a$can be obtained as

$$h^a = LFLB(a_{i,j}), \tag{4}$$

where $a_{i,j}$ is the 3D static feature of the j-th frame in utterance u_i. Similar to the text mode, the conversational context can be represented as $H^a = \{h_1^a, h_2^a, \cdots, h_N^a\}$.

2.3. Intra- modal conversation encoding

The feature extraction process of dialogue feature extraction module in speech and text modes is very similar, so we will not introduce it separately in this section.

Context transformer We use transformers to extract the context features of the session and the original embedding of the conversation, respectively. The utterance to be predicted is treated as a query matrix, and the session context is treated as a key-value matrix. By calculating their self-attention separately, we can know how important each utterance in the context is to the current utterance. Among them, the Q, K, V matrices are obtained by linearization calculation of the context matrix H^* obtained in the last module.

$$[Q^*, K^*, V^*] = Linear([H^*, H^*, H^*]) \tag{5}$$

$$Z^* = softmax(\frac{Q^* K^{\top *}}{\sqrt{d}})V^*, \tag{6}$$

where $* \in \{a, t\}$, represents the speech and text modal.

Speaker-aware transformer Speaker-aware features are also extracted using transformers. Inspired by [20], we use two masks M_p and M_q to calculate intra- and inter-speaker self-attention, respectively. For intra-speaker attention, the mask corresponding to the position of the current speaker's utterances are assigned a value of 1, and the position of the rest of the utterances are assigned a value of negative infinity. Similarly, for inter-speaker attention, the mask corresponding to the position of the current speaker's utterances are assigned a value of negative infinity, and the position of the rest of the utterances are assigned a value of 1. Then the speaker-aware self-attention is calculated as follows

$$Z_p^* = softmax(\frac{Q^* K^{\top *} \odot M_p}{\sqrt{d}})V^* \tag{7}$$

$$Z_q^* = softmax(\frac{Q^* K^{\top *} \odot M_q}{\sqrt{d}})V^*, \tag{8}$$

where $\odot$ means element-wise multiplication.

Through the previous calculation, different self-attention can be obtained, which respectively represent the characteristics of the intra-modal conversation in different aspects and are independent of each other. The above attention heads are concated in series to obtain the multi-head attention in the mode, and then LN and FFN were used for normalization to obtain the output of the model.

$$Z^* = LN(MHA(Z^*) + H^*) \tag{9}$$

$$O^* = max(0, Z^* W_1 + b_1)W_2 + b_2 \tag{10}$$

$$O^* = LN(O^* + Z^*) \tag{11}$$

where LN represents the Layer Normalization, and MHA means multi-head attention.

2.4. Cross-modal feature fusion

In the cross-modal fusion attention module, we modified the attention of the two modes respectively by cross-calculating the dot product between Q and K of the current utterance and context between the modes, so as to obtain the inter-modal influence. Cross-modal attention is calculated as

$$\Delta Z^{a \to t} = softmax(\frac{Q^t K^{\top a}}{\sqrt{d}})V^a, \tag{12}$$

$$\Delta Z^{t \to a} = softmax(\frac{Q^a K^{\top t}}{\sqrt{d}})V^t. \tag{13}$$

Subsequently, $\Delta Z^{a \to t}$ and $\Delta Z^{t \to a}$ are used to update the attention of the original mode to obtain the attention containing cross-modal features.

$$Z^a = LN(Z^a + \Delta Z^{t \to a}) \tag{14}$$

$$Z^t = LN(Z^t + \Delta Z^{a \to t}). \tag{15}$$

Finally, we send the concat of two intra- modal attention and two cross-modal attention to the classifier.

2.5. Emotion classification

For the final classification of emotions, we use a full connection layer and a softmax layer.

$$f_c = tanh(W_f c + b_f) \tag{16}$$

$$o_t = softmax(W_o f_c + b_o) \tag{17}$$

$$\hat{y_t} = \underset{i}{argmax}(o_t[i]), \tag{18}$$

where $i \in [1, k]$, k is the number of emotion classes, and $\hat{y_t}$ is the emotion class predicted by the model. The training loss is as follows

$$\mathcal{L} = -\sum_i y_i log(\hat{y_i}). \tag{19}$$

3. Experiments and Results

3.1. Datasets

We compare the performance of our model with that of the baseline model on the IEMOCAP[23] and MELD[24] dataset. **IEMOCAP** is a multi-modal conversation dataset, with each dialogue performed by two actors based on the script. It consisted of 151 dialogues with 7,433 utterances, each of which was labeled as one of six emotional labels: happy, sad, angry, excited, frustrated, or neutral. **MELD** is also a multi-modal conversation from the TV series Friends. It contains 1433 dialogues and 13,708 utterances, involving 7 emotional categories including anger, disgust, sadness, joy, neutral, surprise, and fear. The division of datasets is shown in Table 1.

Table 1: *The division of IEMOCAP and MELD.*

Dataset	Division	Utterances	Dialogues
IEMOCAP	train	5,810	120
	val		12
	test	1,623	31
MELD	train	9,989	1,039
	val	1,109	114
	test	2,610	280

3.2. Implementation details

Our model is implemented on the PyTorch platform. In text mode, a 300-dimensional Glove is used as word embedding. And utterance features are set to 300 dimensions. In speech mode, we use the openEAR toolkit to extract utterance-level speech features. CNNs utilized in LFLB use convolution kernels of size 3×3, and the step size is (1,1). The initial learning rate is 0.0005, and the decay rate of the learning rate is 0.001. We use the AdamW[25] optimizer. The maximum period of model training was 100 epochs, and the training was stopped if the loss did not decrease for 10 consecutive epochs.

3.3. Baselines

To evaluate the performance of the proposed model, we compared it with the following baselines:

c-LSTM [26] is a context-related discourse representation constructed according to context. C-lstm+att uses attention at each timestep, making it better able to focus on important contextual information. **DialogueRNN** [18] distinguishes different participants in a conversation in the form of interaction, and models the context, state, and emotion of different speakers. **IEIN** [27] is a simulation of a speaker's emotional interaction using emotional labels based on the iterative prediction of conversation progression. **BiERU** [28] applies a two-way emotional circular unit that takes full advantage of both parties in

Table 2: *Performance comparison of the 10-fold cross-validation in terms of Accurancy and F1 between the proposed SCFA with the baselines on the IEMOCAP and MELD. * indicates that the improvement achieved by our model compared to the optimal baseline is statistically significant under the 1-tailed t-test ($p < 0.05$).*

Method	IEMOCAP														MELD	
	happy		sad		neutral		angry		excited		frustrated		**Average**		**Average**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	wF1	Acc	wF1
c-LSTM+att	30.56	35.65	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	**65.88**	59.41	56.32	56.19	57.50	55.71
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	65.28	**80.27**	71.86	61.15	58.91	63.40	62.75	56.10	55.90
IEIN	-	**53.17**	-	77.19	-	**61.31**	-	61.45	-	69.23	-	60.92	-	64.37	-	60.72
BiERU	**49.81**	32.75	81.26	**82.37**	5.00	60.45	67.86	65.39	63.14	73.29	59.77	60.68	65.35	64.24	60.70	60.48
TRMSM	43.36	50.22	81.23	75.82	**66.11**	64.15	60.39	60.97	77.64	72.70	62.16	63.45	65.34	65.74	63.23	62.36
MM-DFN	-	-	-	-	-	-	-	-	-	-	-	-	-	65.41	-	58.34
SCFA(Ours)	46.24	47.61	**82.23**	81.30	63.77	59.08	**70.13***	**68.52***	74.26	**74.63***	64.38	**64.29***	**67.91***	**66.42***	**64.86***	**63.69***

Table 3: *Ablation study on the IEMOCAP database. In the table, T and A represent text and audio modal input only, and w/o CFA, w/o CTr, and w/o SaTr represent the performance without the proposed Cross-modal Fusion Attention, Context Transformer, and Speaker-aware Transformer, respectively.*

Approach	Acc	wF1
SCFA	67.91	66.42
T	63.82	62.89
A	49.24	48.66
w/o CFA	62.44	62.88
w/o CTr	64.35	62.93
w/o SaTr	65.65	65.37

Table 4: *Ablation study on the MELD database.*

Approach	Acc	wF1
SCFA	64.86	63.69
T	59.32	57.76
A	47.37	44.18
w/o CFA	61.24	59.70
w/o CTr	62.09	61.43
w/o SaTr	61.33	60.21

the context of the conversation to predict emotions, regardless of the parties involved. **TRMSM**[20] is a hierarchical approach that utilizes three masks to model speaker dependencies. **MM-DFN**[29] designs a graph-based dynamic fusion module for fusing multi-modal context features in a conversation.

3.4. Experimental results

We used Accuracy and F1 to evaluate the performance of the models on both datasets, and the experimental results are shown in Table 2. On the IEMOCAP data set, our model achieve the best results in 8 out of all 14 indicators and achieved the best average ACC and F1. And nn the MELD dataset, SCFA improves by 4.1 and 2.8 points on ACC and F1, respectively, compared with the previous optimal results. By observing the performance of different emotion categories, SCFA has a more balanced performance compared to the baselines, indicating that it can better identify various emotions. It is worth noting that SCFA shows the best performance in the more intense emotions such as sad, angry, and excited. The possible reason is that when the speaker expresses these emotions, the emotional features contained in the voice are more obvious, which is easier to distinguish from the relatively gentle emotions. The above results demonstrate the superiority of the SCFA.

3.5. Ablation study

In order to explore the role of each part of our model, we conducted an ablation study. The results on the two datasets are shown in Table 3 and Table 4, where T and A represent the results of removing the speech modal and text modal based on our model, respectively. W/o CFA represents to removal of the CFA from the model and only provides the features of two single modes to the classifier. W/o CTr and w/o SaTr denote the removal of the context transformer and the speaker-aware transformer in the intra-modal conversation encoding module, respectively.

Obviously, compared with the single-mode method, the cross-modal method has a fair improvement in performance. This indicates that different modes contain complementary emotional information. In the absence of cross-modal fusion, the result of ACC and wF1 decreased by more than 3 in both datasets. The experimental results verify the effectiveness of the CFA. It can be seen from the experimental results that removing the context transformer module also has a significant impact on Acc and wF1. As for the speaker-aware transformer module, the ablation results of the two datasets seem to show different effects. The speaker-aware transformer appears to have little effect on model performance on IEMOCAP, while more pronounced on MELD. A possible explanation for this result is that there are only two speakers per conversation in IEMOCAP, and MELD is a multi-party conversation dataset, so our model can better capture speaker interactions on MELD.

4. Conclusions

In this paper, we propose a speaker-aware cross-modal fusion attention model (SCFA) for conversational emotion recognition. For the conversation context, the model combines the original embeddings with the conversation transformer to roundly preserve the context characteristics. For speaker-aware features, the model obtains the intra- and inter-speaker influence through different masks. The conversation encoding within speech and text modal are fetched by utilizing the above two mechanisms. Furthermore, we introduce CFA to capture the inter-modal interactions. The average performance on both the IEMOCAP and MELD datasets has improved to some extent compared with baselines. And the results of ablation studies also show the validity of the various parts of the proposed model.

5. Acknowledgements

This work was supported by the National Science Foundation of China under Grant 62076092 and Special Project of Foshan Science and the Technology Innovation Team under Grant FS0AA-KJ919-4402-0069.

6. References

[1] C.-C. Hsu and L.-W. Ku, "Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues," in *SocialNLP*, 2018, pp. 27–31.

[2] W. Jiao, H. Yang, I. King, and M. R. Lyu, "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition," *arXiv preprint arXiv:1904.04446*, 2019.

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

[4] D. Hu, L. Wei, and X. Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 7042–7052. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.547

[5] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "A dialogical emotion decoder for speech emotion recognition in spoken dialog," in *ICASSP*, 2020, pp. 6479–6483.

[6] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[7] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku *et al.*, "Emotionlines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379*, 2018.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *SLT*, 2018, pp. 112–118.

[9] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint arXiv:1911.00432*, 2019.

[10] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018, pp. 2247–2256. [Online]. Available: https://aclanthology.org/P18-1209/

[11] Y. Fu, S. Okada, L. Wang, L. Guo, Y. Song, J. Liu, and J. Dang, "CONSK-GCN: conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition," in *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*. IEEE, 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME51207.2021.9428438

[12] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *EMNLP*, 2018, pp. 3454–3466.

[13] E. Kim and J. W. Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *ICASSP*, 2019, pp. 6720–6724. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683077

[14] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychological science: a journal of the American Psychological Society*, no. 7, p. 21, 2010.

[15] S. Poria, N. Majumder, R. Mihalcea, and E. H. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2929050

[16] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2018, pp. 2594–2604. [Online]. Available: https://doi.org/10.18653/v1/d18-1280

[17] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," 2018, pp. 2122–2132. [Online]. Available: https://doi.org/10.18653/v1/n18-1193

[18] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *AAAI*, no. 01, 2019, pp. 6818–6825.

[19] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. F. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1015

[20] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, "A hierarchical transformer with speaker modeling for emotion recognition in conversation," *CoRR*, vol. abs/2012.14781, 2020. [Online]. Available: https://arxiv.org/abs/2012.14781

[21] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018. [Online]. Available: https://doi.org/10.1109/LSP.2018.2860246

[22] H. Zhao, Y. Gao, and Y. Xiao, "Upgraded attention-based local feature learning block for speech emotion recognition," in *PAKDD*, 2021, pp. 118–130.

[23] C. Busso, M. Bulut, 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. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[24] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2019, pp. 527–536. [Online]. Available: https://doi.org/10.18653/v1/p19-1050

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[26] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Association for Computational Linguistics, 2017, pp. 873–883. [Online]. Available: https://doi.org/10.18653/v1/P17-1081

[27] X. Lu, Y. Zhao, Y. Wu, Y. Tian, H. Chen, and B. Qin, "An iterative emotion interaction network for emotion recognition in conversations," D. Scott, N. Bel, and C. Zong, Eds., 2020, pp. 4078–4088. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.360

[28] W. Li, W. Shao, S. Ji, and E. Cambria, "Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2021.09.057

[29] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 7037–7041. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747397