MULTI-STAGE GRAPH REPRESENTATION LEARNING FOR DIALOGUE-LEVEL SPEECH EMOTION RECOGNITION

Yaodong Song[1,†], *Jiaxing Liu*[1,†], *Longbiao Wang*[1,*], *Ruiguo Yu*[1,*], *Jianwu Dang*[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{songyaodong, jiaxingliu, longbiao_wang, rgyu}@tju.edu.cn, jdang@jaist.ac.jp

ABSTRACT

With the development of speech emotion recognition (SER), most of current research is utterance-level and cannot fit the need of actual scenarios. In this paper, we propose a novel strategy that focuses on capturing dialogue-level contextual information. On the basis of utterance-level representation learned by convolutional neural network (CNN) which is followed by the bidirectional long short-term memory network (BLSTM), the proposed dialogue-level method consists of two modules. The first module is Dialogue Multi-stage Graph Representation Learning Algorithm (DialogMSG). The multi-stage graph that modeling from different dialogue scope is introduced to capture more effective information. The other one is a double-constrained module. This module includes not only an utterance-level classifier but also a dialogue-level graph classifier which is named as Atmosphere. The results of extensive experiments show that the proposed method outperforms the current state of the art on the IEMOCAP benchmark dataset.

Index Terms— Speech emotion recognition, dialogue-level contextual information, utterance-level representation, double-constrained, atmosphere

1. INTRODUCTION

Speech Emotion Recognition (SER) has attracted more and more attention from researchers , because it not only has great application prospects in human-computer interaction, but also has increasingly wide applications in the fields of intelligent cognitive assistant [1], online education [2], computer games [3] and so on.

In recent years, some deep learning methods such as the Deep Neural Networks (DNNs) [4], Convolutional Neural Networks (CNNs) [5], Recurrect Neural Networks (RNNs) [6, 7] and Attention mechanism [8, 9] have been used in the SER task and achieved competitive results. In particular, Satt et al. [10] proposed the CNN-BLSTM method was a classical work, which firstly introduced CNN to extract features, and then the bidirectional long short-term memory network (BLSTM) [11] was employed to learn intra-utterance contextual information. Although the current research had achieved good results, there was still a gap to deploy them to dialogue scenarios in actual applications. Recently, some natural language processing (NLP) works tried one step further, such as DialogueRNN [12]. The method introduced RNN with attention mechanism to learn inter-utterance information from the most relevant text of future and past. However, the performance of long-range contextual modeling was subject to the structure of RNN. Another representative work was the introduction of graph convolutional neural network (GCN) [13], such as DialogueGCN [14] and RGAT [15]. Both of these works introduced the GCN to capture the relations of utterances in one dialogue and also achieved competitive result in NLP. Due to the over-smoothing problems existing in GCNs, the number of convolutional layers could not flexibly cope with different ranges. At the same time, the above-mentioned methods were only modeling one sentence, and lacked the consideration of the whole current dialogue.

To handle the problems mentioned above, we propose a novel dialogue-level contextual information learning system as shown in Fig. 1. Firstly, we introduce the CNN-BLSTM to extract utterance-level representation from the spectrogram. Secondly, the learned utterance-level representation are converted to dialogue-level format. The following dialogue-level strategy consists of two modules. In the first module, the Dialogue Multi-stage Graph Representation Learning Algorithm (DialogMSG) is proposed to capture the context dependence in each dialogue. In the graph building, the utterances in dialogue are represented as nodes, and the dependence relations of the utterances are regarded as edges. Considering the sequence modeling, future information should not be leaked to the past. A unidirectional Gate Recurrent Unit (GRU) [16] based Local-wised Update Mechanism is employed to represent past contextual information. To further capture long-range contextual information, a multi-stage

[*] CORRESPONDING AUTHOR

[†] EQUAL CONTRIBUTION

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

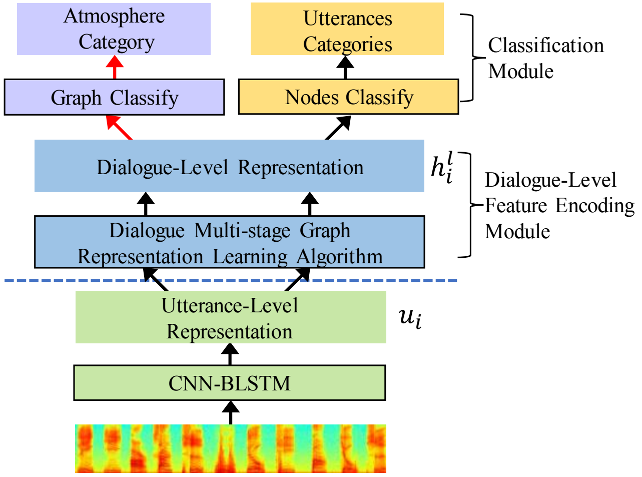
structure integrating residual connections is also brought in. The residual multi-stage GRU-based graph structure ensures that the key information can still be effectively retained even in a long-range dialogue. Moreover, the problems of over-smoothing in DialogueGCN can also be well alleviated. In the second module, traditional classification constraint of each utterance in one dialogue is set. At the same time, A constraint on the classification of dialogue graphs is also taken into account which is named as Atmosphere.

In summary, our contributions are as follows:

- The DialogMSG is proposed to capture contextual information from local region to global region through residual multi-stage GRU-based graph.
- For the first time, a constraint named as Atmosphere is taken into account to get a higher sensitivity to the changes of emotion in the long-rang dialogues.

2. PROPOSED METHOD

The proposed strategy in this paper is shown in Fig.1. As shown in Fig.1, we introduce the CNN-BLSTM model as the utterance-level contextual information learned method. Formally, the dataset consists of M dialogues, one dialogue consists of N utterances $u_1, u_2, \dots, u_N$. Also, $u_t \in \mathbb{R}^{D_m}$ denotes utterance-level representation. On the basis of utterance-level representation learned, we propose a dialogue-level innovation strategy. The details are introduced below.



Atmosphere Category
Utterances Categories
Classification Module
Graph Classify
Nodes Classify
Dialogue-Level Representation
h_i^l
Dialogue-Level Feature Encoding Module
Dialogue Multi-stage Graph Representation Learning Algorithm
Utterance-Level Representation
u_i
CNN-BLSTM

Fig. 1. Our proposed dialogue-level SER architecture; red arrow denotes only appearing in the training stage.

2.1. Dialogue-Level Feature Encoding Module

Since, multi-party dialogue is a natural graph structure, we propose the Dialogue-Level Feature Encoding Module in the form of graph to capture inter-utterances contextual information in a dialogue. Fig.2 shows the details of this module.

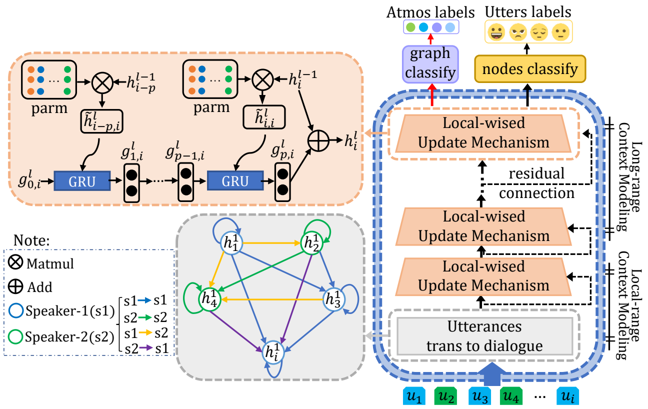


Fig. 2. The details of DialogMSG.

Some notations will be introduced: a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$ is construted from a dialogue having N utterances, which nodes(entities) $h_i \in \mathcal{V}$, directed edges $e_{j,i} \in \mathcal{E}$.Meanwhile, relation $r \in \mathcal{R}$ is the type of edge between h_i and h_j and $\delta_{j,i} \in [0, 1]$ is the weight of directed edges $e_{j,i}$, where $\delta_{j,i} \in \mathcal{W}$ and $i, j \in [1, 2, \dots, N]$.

First, we transform the utterance-level feature representation u_i to dialogue format and construct a graph structure. Each utterance in a dialogue is represented as a node $h_i \in \mathcal{V}$ in the graph $\mathcal{G}$, where the node representation h_i is initialized with u_i for all $i \in [1, 2, \dots, N]$. Considering the natural dialogue scene in reality and model complexity, we construct the edges by only considering the past context window size of p and no considering any future utterances information. Hence, each node in the graph is connected to at most $p + 1$ nodes (including itself) with an edge.

The edge weights in the graph is formulated as:

$$\delta_{j,i} = \text{softmax}\left(h_i^T W_s h_j\right) \tag{1}$$

$$\text{softmax}(x) = [e^{x_1}/\Sigma_i e^{x_i}, e^{x_2}/\Sigma_i e^{x_i}, \dots] \tag{2}$$

where $j = i-p, \dots, i$ and W_s is a trainable parameter matrix. In Eq.(1) and Eq.(2), we calculate the edge weights based attention module. It will assign higher attention scores to the utterances emotionally relevant to h_i.

Then, we feed the utterance-level representation h_i and edge weights $\delta_{j,i}$ into the Local-wised Update Mechanism:

$$\widetilde{h_{j,i}}^{(1)} = \frac{\delta_{j,i}}{q_{i,r}} W_r^{(1)} h_j \tag{3}$$

$$g_{k,i}^{(1)} = \overrightarrow{GRU_{\mathcal{S}}}\left(g_{k-1,i}^{(1)}, \widetilde{h_{j,i}}^{(1)}\right) \quad \text{for } j = i-p, \dots, i; k = 1, \dots, p \tag{4}$$

$$h_i^{(1)} = g_{p,i}^{(1)} + W_o^{(1)} h_i \tag{5}$$

where relation $r \in \mathcal{R}$, $\mathcal{R}$ only contains relations in canonical direction(e.g. *born_in*) and no contains in inverse direction(e.g. *born_in_inv*) due to only considering past contextual information. $q_{i,r}$ is a problem-specific normalization constant that can be learned and $W_r^{(1)}$, $W_o^{(1)}$ are learnable parameters of the model. Eq.(3)-(5) are the update mechanism of the first stage of this module. We consider the speaker information of each party in Eq.(3). In theory, RNN and GRU can propagate long-range contextual information well. But it doesn't always work well in practice. So we use unidirectional GRU cell $\overrightarrow{GRU_S}$ to capture local-range (at most past p utterances) contextual information in Eq.(4). Our aim is to make the model encode local-range context representation more efficient and focused due to less emotional dynamics. Since this model is multi-stage, we use a structure similar to residual connection in Eq.(5).

In order to make the model have the ability to capture long-range contextual information, we carry out multi-stage modeling:

$$g_{k,i}^{(l)} = \overrightarrow{GRU_{\mathcal{L}}}\left(g_{k-1,i}^{(l)}, W^{(l)} h_j^{(l-1)}\right), \quad \text{for } j = i-p, \ldots, i; k = 1, \ldots, p \tag{6}$$

$$h_i^{(l)} = g_{p,i}^{(l)} + W_o^{(l)} h_i^{(l-1)} \tag{7}$$

where $l \geq 2$, $W^{(l)}$ and $W_o^{(l)}$ are learnable parameters. In this paper, we set $l = 6$. In the sixth layer of the model, the model can take into account the information of the first 61 utterances to some extent which can completely cover the average length of the dialogue.

2.2. Classification Module

The module contains two classifiers: graph classifier and nodes classifier. Here, we propose a new concept: dialogue atmosphere. We think the atmosphere of the whole dialogue is closely related to the emotion of each speaker in a dialogue. So we consider the graph modeling for the whole dialogue to constrain the learning process. We take the emotion that appears most in a dialogue as an atmosphere label of the whole dialogue for graph classification. To the authors' best knowledge, this is the first time in the literature to consider the factor of dialogue atmosphere in the task of emotion recognition.

The dialogue-level feature representation h_i^l is learned by previous mentioned module. For graph classifier, we use a readout layer [17] that aggregates node features to make a fixed size graph representation.

$$l_i = \frac{1}{N} \sum_{i=1}^{N} h_i^l \| \max_{i=1}^{N} h_i^l \tag{8}$$

where N is the number of nodes, h_i^l is the feature vector of i-th node, and $\|$ denotes concatenation.

For nodes classifier, the h_i^l is followed by a similarity-based attention mechanism to get a final nodes representation $\tilde{h}_i$.

$$\theta_i = \operatorname{softmax}\left((h_i^l)^T W_\beta \left[h_1^l, h_2^l \ldots, h_N^l\right]\right) \tag{9}$$

$$\tilde{h}_i = \theta_i \left[h_1^l, h_2^l, \ldots, h_N^l\right]^T \tag{10}$$

Finally, the l_i and $\tilde{h}_i$ are fed to a fully-connected network and $softmax$ layer to get the dialogue atmosphere and utterances classfication results, respectively.

3. EXPERIMENTS AND ANALYSIS

3.1. Experimental Setup

We evaluate our method on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [18] database. Following reported procedure in prior works, we only use 5531 audio utterances in 151 dialogues with four emotion categories: Neutral (30.88%), Anger (19.94%), Sad (19.60%) and Happy (29.58%). Since the length of each segment containing effective emotional information is an open question, we use the same preprocessing method as guo et al. [19]. The time of each segment is 265-ms and the input spectrogram has the following $time \times frequency$: 32×129. We use the sessions 1-4 as the training set and use session 5 as test set. We choose cross-entropy as the cost function both in graph classify and nodes classify and Adam as the optimizer. The final loss of the model is the sum of the two losses. The window size of the past is set as 10.

3.2. Experiment Results and Analysis

To verify the effectiveness of the proposed method, we set up three groups of experiments. The first set of experiment is visual analysis to illustrate the effectiveness of the dialogue-level strategy. The second group is to show the classification results of comparative experiments. The weighted accuracy (WA), unweighted accuracy (UA) and f1-score (F1) are employed for performances assessment. The last group is to illustrate the effectiveness of the dialogue atmosphere.

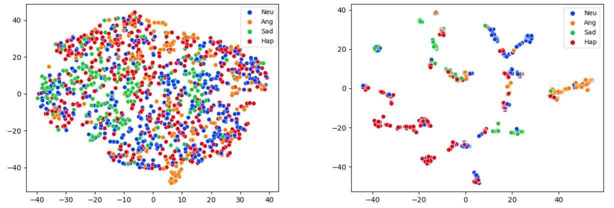
3.2.1. Visual analysis

To observe extracted the utterance-level representation and dialogue-level representation, we introduce the t-distributed Stochastic Neighbor Embedding (t-SNE) [20] to visualize the four emotional categories as shown in Fig.3.

Fig.3(a) and Fig.3(b) show the visualization result of utterance-level representation (i.e. the DialogMSG input) and dialogue-level representation (i.e. the DialogMSG output), respectively. It can be observed that the visualization result of the former is very terrible, but the features after the DialogMSG are projected into different clusters and are separated from each other. The also illustrates the effectiveness of our proposed the dialogue-level innovation strategy.

3.2.2. Comparative experiments

For a comprehensive evaluation, we compare the performance of our model with the current advanced approaches, as shown



(a) utterance-level representation (b) dialogue-level representation

Fig. 3. The t-SNE visualizations of extracted representation.

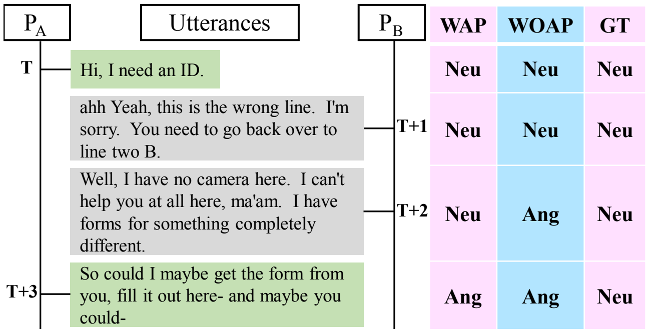
in Table 1. From this table, we observe that, the method that considers the dialogue atmosphere are more effective than the method that do not, demonstrating the significance of it. And our model is around 1.29% (WA), 2.53% (UA) and 4.26% (F1) better than the state-of-the-arts with dialogue-level representation and at least 12.66% better than other baseline model with utterance-level representation. The further illustrates the effectiveness of our model in the dialogue-level SER task.

Table 1. The results of comparative experiments. † with utterance-level representation; ‡ with dialogue-level representation; atmos stands for the word atmosphere.

Models	WA(%)	UA(%)	F1(%)
CNN-BLSTM [10]†	53.83	57.10	53.23
DialogueRNN [12]‡	66.32	67.23	65.70
DialogueGCN [14]‡	68.90	66.68	66.22
DialogMSG w/o atmos	69.14	69.47	69.38
DialogMSG (Ours)	**70.19**	**69.76**	**70.48**

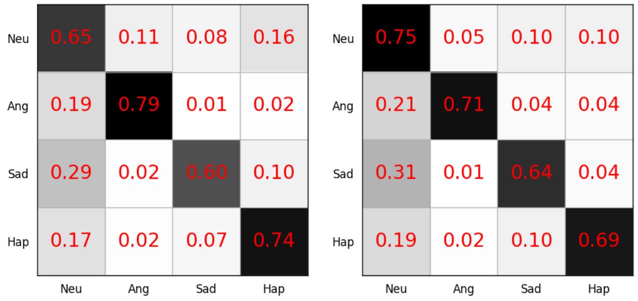
3.2.3. *Effectiveness of dialogue atmosphere*

To further verify the effectiveness of the dialogue atmosphere and the performance of the proposed model, we analyze cases that are correctly predicted from model with dialogue atmosphere (referred to as WA) but incorrectly predicted from model without atmosphere (referred to as WOA). Fig.4 shows a dialogue case. We observe these two models are predicted inconsistently at time $t + 2$. It may be that the WA has perceived that the overall atmosphere is biased towards neutrality at the beginning of the dialogue lead to correct prediction. However, both of them are wrong at time $t + 3$. This is because at this moment, the text is neutral, but the voice sent by P_A is a little angry. And, by analyzing these cases, we also observe that when there are only two emotional categories in a conversation, the WOA tends to predict all the utterances of the conversation as the same category, while the WA has an interaction of the two emotions. From Fig.5, the phenomena can be observed. Fig.5(a) shows that anger and sad emotion exhibit two extremes, while in Fig.5(b) it is shown that this phenomenon tends to moderate. This also means that the interaction through emotional can reduce the false positive rate of the model to some extent and improve the true positive rate of the model.



P_A	Utterances	P_B	WAP	WOAP	GT
T	Hi, I need an ID.		Neu	Neu	Neu
	ahh Yeah, this is the wrong line. I'm sorry. You need to go back over to line two B.	T+1	Neu	Neu	Neu
	Well, I have no camera here. I can't help you at all here, ma'am. I have forms for something completely different.	T+2	Neu	Ang	Neu
T+3	So could I maybe get the form from you, fill it out here- and maybe you could-		Ang	Ang	Neu

Fig. 4. Conversation cases from IEMOCAP session5 with WAP (With-Atmosphere-Predicted), WOAP (WithOut-Atmosphere-Predicted) and GT (Ground-Truth) emotions.



	Neu	Ang	Sad	Hap
Neu	0.65	0.11	0.08	0.16
Ang	0.19	0.79	0.01	0.02
Sad	0.29	0.02	0.60	0.10
Hap	0.17	0.02	0.07	0.74

	Neu	Ang	Sad	Hap
Neu	0.75	0.05	0.10	0.10
Ang	0.21	0.71	0.04	0.04
Sad	0.31	0.01	0.64	0.04
Hap	0.19	0.02	0.10	0.69

(a) DialogMSG (w/o atmos) (b) DialogMSG (w/ atmos)

Fig. 5. One group of confusion matrices of proposed method.

4. CONCLUSION

In this paper, we proposed a dialogue-level SER framework. On the basis of utterance-level representation learned by CNN-BLSTM, we proposed dialogue-level method which consists of two modules. The first module is the Dialogue Multi-stage Graph Representation Learning Algorithm(DialogMSG) to extract dialogue-level contextual information. The second is a double-constrained module where we consider the factor of dialogue atmosphere to assist the model in the classifying utterances. These components work together improve the model performance. Experimental results on the IEMOCAP dataset show that our model achieves the state-of-the-art performance, and extensive analysis further proves the effectiveness of our model.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200 and the National Natural Science Foundation of China under Grant 62176182.

6. REFERENCES

[1] Alexandra König, Linda E Francis, Aarti Malhotra, and Jesse Hoey, "Defining affective identities in elderly nursing home residents for the design of an emotionally intelligent cognitive assistant," in *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2016, pp. 206–210.

[2] Oussama El Hammoumi, Fatimaezzahra Benmarrakchi, Nihal Ouherrou, Jamal El Kafi, and Ali El Hore, "Emotion recognition in e-learning systems," in *International Conference on Multimedia Computing and Systems(ICMCS)*, 2018, pp. 1–6.

[3] Daniel Leite, Volnei Frigeri Jr, and Rodrigo Medeiros, "Adaptive gaussian fuzzy classifier for real-time emotion recognition in computer games," *arXiv preprint arXiv:2103.03488*, 2021.

[4] E. Kim and J. W. Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6720–6724, 2019.

[5] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang, "Time-frequency deep representation learning for speech emotion recognition integrating self-attention," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 681–689.

[6] Amr Mostafa, Mahmoud I. Khalil, and Hazem Abbas, "Emotion recognition by facial features using recurrent neural networks," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 2018, pp. 417–422.

[7] Yunfeng Xu, Hua Xu, and Jiyun Zou, "Hgfm: A hierarchical grained and feature model for acoustic emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6499–6503.

[8] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, pp. 114683, 2021.

[9] Srividya Tirunellai Rajamani, Kumar T Rajamani, Adria Mallol-Ragolta, Shuo Liu, and Björn Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6294–6298.

[10] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms.," in *Interspeech*, 2017, pp. 1089–1093.

[11] Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[12] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6818–6825.

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, pp. 3844–3852, 2016.

[14] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," 2020, pp. 154–164, EMNLP-IJCNLP.

[15] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7360–7370.

[16] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò, "Towards sparse hierarchical graph classifiers," *arXiv preprint arXiv:1811.01287*, 2018.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[19] Lili Guo, Longbiao Wang, Jianwu Dang, Linjuan Zhang, and Haotian Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2666–2670.

[20] Laurens Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*, 2009, pp. 384–391.