International Journal of Engineering Research And Management (IJERM)
ISSN: 2349- 2058, Volume-10, Issue-11, November 2023

Directed Acyclic Graphs using Approximate Labels for Conversational Emotion Recognition

Wenxue Zhang, Baoshan Sun

Abstract— **Conversational context modelling is a crucial aspect that holds significance in emotion recognition from conversation. In this paper, we present the directed acyclic graph using approximate labels model(LDAG-ERC), which aims to enhance the performance of the current ERC dialogue task model, DAG-ERC.The LDAG-ERC model is a modified version of the DAG-ERC model. It uses directed acyclic graphs (DAG) with approximate emotion labels to describe the internal structure of the dialogue. Our LDAG-ERC model has significantly improved performance compared to the DAG-ERC model. We conducted experiments on an ERC datasets and compared them to the original model to illustrate the effectiveness of the modifications.**

Index Terms—**DAG,ERC,NLP.**

I. Introduction

The objective of the ERC task is to analyse the emotions conveyed in each sentence of a conversation. By examining the emotional state of each speaker, the ERC task can assist in social media opinion analysis, dialogue system evaluation, and the creation of empathy machines within various fields, including healthcare.

However, in practice, emotional responses to dialogue are influenced by numerous uncontrollable factors and can fluctuate, making it challenging to identify emotions in the ERC task. Additionally, the current conversational context plays a significant role in eliciting emotions, resulting in many studies seeking to enhance conversational context modelling.

Methods for modelling context in ERC tasks are broadly divided into two categories: graph-based and recurrence-based. Graph-based methods typically gather data related to the surrounding utterances while disregarding information about more distant utterances and the order in which they are produced. Recursion-based methods consider the remote discourse and sequence information in comparison to graph-based methods. However, the information employed to update the status of the query statement is insufficient, and the information at the distant end is weakened layer by layer during propagation, significantly hampering their effectiveness.

Dialogues possess the subsequent features: (1) dialogues do not have the ability to anticipate forthcoming information, they can only obtain previous information, (2) each interlocutor exerts an influence on the others involved in the conversation, (3) each interlocutor contemplates on the statements they articulated[1], and (4) a dialogue can consist of more than one sentiment[2]. In summary, it can be inferred that directed acyclic graphs (DAG) satisfy the previous two requirements and combine the benefits of graph and recursive approaches. Therefore, we chose to improve on the DAG-ERC[3] model.The remaining two requirements outlined in the paper can be attained through the modifications made.

We conducted experiments on an ERC dataset to assess the ensuing performance improvement, and the results of the experiments affirm the effectiveness of the changes.

II. Related Work

A. Emotion Recognition in Conversation

Emotion recognition in dialogue is the task of identifying emotions by classifying a piece of dialogue. Unlike emotion recognition in single sentences, emotion recognition in dialogue is more complex because it requires consideration of several factors, including the discourse itself and its context, the state of the speaker, and the emotion expressed in the preceding discourse.

B. Recent Studies

Today's ERC models comprise of DialogueCRN for deep information extraction during reasoning, recursion-based DialogRNN[4], CoMPM[5] for modelling through the extraction of information from pre-trained models, graph-based DialogGCN[6], and DualGATs[7] for modelling speaker and utterance relationships.

III. Methodology

A. Problem Definition

A conversation can be seen as a sequence of successive utterances, and the conversation S_i is defined as $\{u_1, u_2, ..., u_N\}$. N represents the number of utterances in the conversation, where $u_i = \{w_{i1}, w_{i2}, ..., w_{in_i}\}$, n_i denotes the number of tokens in the utterance u_i , w_i denotes the token in the utterance u_i , $Y_i = \{y_1, y_2, ..., y_N\}$ denotes the emotion label of each utterance in the conversation S_i , and the speaker is defined as a function $p(\cdot)$, the speaker of u_i is represented by $p(u_i)$.

Manuscript received November 13, 2023

Wenxue Zhang, School of computer science and technology, Tiangong university, Tianjin ,China
Baoshan Sun, School of computer science and technology, Tiangong university, Tianjin ,China

B. LDAG-ERC Layers

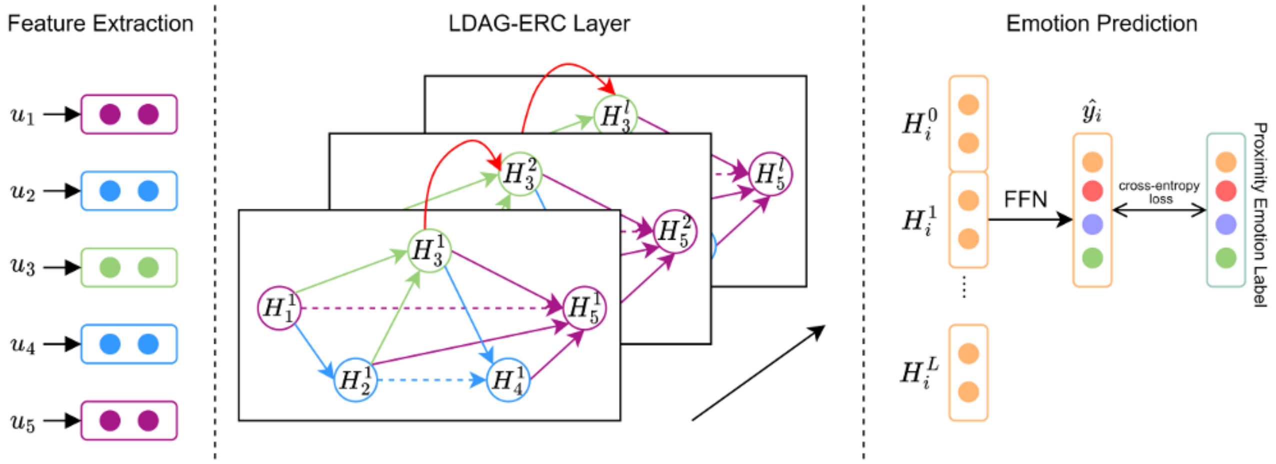

Figure 1. The frame work of LDAG-ERC

We believe the speaker will reflect on current utterances as he/she speaks, so DAG-ERC's algorithm for constructing directed acyclic graphs is modified so that each node is connected to its own node in the previous layer, representing the speaker's reflection on current utterances, as seen in Fig.1.

For each utterance ui,the attentional weights are computed with respect to its own predecessor node and its own hidden state in the $(l-1)$ -th layer.

$$\alpha_{ij}^{l} = \text{Softmax}_{j \in N_i}(W_{\alpha}^{l}[(H_j^{l} \,\|\, H_i^{l-1}) \oplus (H_i^{l-1} \,\|\, H_i^{l-1})]) \quad (1)$$

Where W_α^l are trainable parameters, $\|$ denotes the concatenation operation, it signifies the combination of feature vectors, and $\oplus$ denotes the parallel connection . N_i denotes the predecessor nodes of the i -th node.

The construction of the directed acyclic graph follows the algorithm of DAG-ERC.

Table 1. The emotion vector, where valence is the x-axis and arousal is the y-axis.

	Valence	Arousal
neutral	0	0
happy	cos(3*π/20)	sin(3*π/20)
sad	−cos(9*π/20)	−sin(9*π/20)
angry	−cos(9*π/20)	sin(9*π/20)
frustrated	−cos(π/20)	−sin(π/20)
excited	cos(7*π/20)	sin(7*π/20)

C. Training and Prediction

For emotions that are similar to utterance emotion label, we allow a small fraction of the emotion label to belong to the emotion label similar to it and gradually normalise it to a one-hot encoding in the following iterations. Initially, we compute the similarity between emotions to derive the similarity matrix M_{sim}. The computation of the similarity matrix follows the following formula.

$$s_{i,j} = \begin{cases} \max(\cos(v_i, v_j), 0) & v_i \cdot v_j > 0 \\ 0 & v_i \cdot v_j < 0 \\ 1/N & v_i \cdot v_j = 0 \end{cases} \quad (2)$$

where v denotes the emotion vector and the emotion vector is given by Table 1[8]. After obtaining the similarity matrix, the proximity emotion label are normalised to the [0,1] range.

At the end, The output of all the layers will be stitched together as the final representation of the u_i and predict the emotion through the feed-forward neural network. We use the cross-entropy loss as the objective function during the training process.

$$H_i = \|_{l=0}^{L} H_i^l \quad (3)$$

$$z_i = \text{ReLU}(W_H H_i + b_H) \quad (4)$$

$$P_i = \text{Softmax}(W_z z_i + b_z) \quad (5)$$

$$\hat{y}_i = \text{Argmax}_{k \in S}(P_i[k]) \quad (6)$$

$$L(\theta) = -\sum_{i=1}^{M}\sum_{t=1}^{N_i} \text{Log}P_{i,t}[\tilde{y}_{i,t}] \quad (7)$$

M is the number of conversations in the training set, N_i is the number of the utterance in the i-th conversation, $\tilde{y}_{i,t}$ is the proximity emotion label, from the normalised similarity matrix M_{sim}.

Table 2. IEMOCAP dataset

Dataset	Conversations			Urterrances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623

IV. Experiment

A. Dataset

We evaluated the performance of LDAG-ERC using the IEMOCAP[9] dataset.

IEMOCAP The dataset was used to record facial, head and hand markings from 10 actors, providing detailed information about their facial expressions and hand movements during scripted and spontaneous spoken communication scenes. The actors performed selected emotional scripts and improvised hypothetical scenes designed to elicit specific types of emotions. The IEMOCAP dataset contains approximately 12 hours of audiovisual data, including video, voice, facial motion capture, and text transcription. The IEMOCAP dataset categorises emotions into six emotions: neutral, happy, sad, angry, frustrated, and excited.

The statistics of the IEMOCAP dataset are shown in Table 2.

B. *Experiment Details*

The experiment was carried out utilising the features of the dataset extracted from DAG-ERC. The hidden vectors and dimensions for the other hyperparameters were set to 300. Additionally, a window size of $\omega=1$ was applied. The model was executed for 30 epochs on the IEMOCAP dataset. F1 score and accuracy were considered in evaluating the outcome of the experiment.

C. *Results and Analysis*

Table 3 reports the effect of the LDAG-ERC model on the IEMOCAP dataset in comparison with the original model. By observing the results, we can find that the F1 score of LDAG-ERC is improved by 0.61 and the accuracy is improved by 0.62 compared to DAG-ERC, so LDAG-ERC is better than DAG-ERC. Therefore we can affirm that nodes reflecting on their own information and the existence of similarity between emotions are two important features in conversations. The experimental results show that the changes we have made are effective.

Table 3. Model Performance Comparison

model	F1-score	Accuracy
DAG-ERC	67.92	67.94
LDAG-ERC	**68.53**	**68.56**

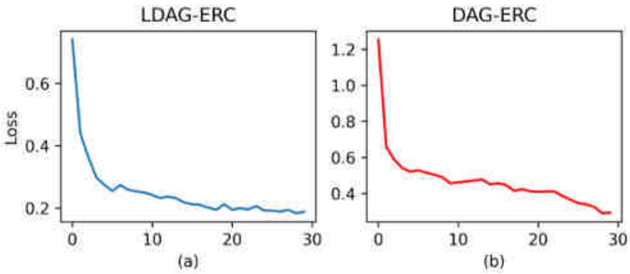

Figure 2. Comparison of LDAG-ERC and DAG-ERC loss plots

Looking at the loss comparison graph Fig.2, it can be seen that the initial loss value of LDAG-ERC is smaller, and the loss decreases slowly as the epoch increases, eventually tending to stabilise. Whereas the initial loss of DAG-ERC is larger and the loss decreases more slowly and fluctuates more during the training process.

D. *Ablation Study*

In order to evaluate the effectiveness of the two changes added to the LDAG-ERC, we removed the two changes separately to observe the change in model performance.

As shown in the Table 4, the performance of the model decreases after removing the two changes, which means that the two changes are valid.

It can be seen that after removing self-reflection, the F1 score of the model decreases by 0.42, indicating that self-reflection plays a larger role in the conversation. After removing proximity labels, the F1 score of the model decreased by 0.14, which proves that we add proximity emotion labels are effective. And it is a performance improvement at a very small cost.

Table 4. Results of ablation experiments on the IEMOCAP dataset

Method	IEMOCAP
LDAG-ERC	68.53
w/o simlabel	68.39(↓0.14)
w/o self-reflection	68.11(↓0.42)

V. Conclusion

This paper presents a novel LDAG-ERC model achieved through modifications to the DAG-ERC model and iterative sentiment label normalisation. Our experimental results demonstrate that our amendments significantly enhance the model performance at minimal cost. Moreover, our experiments suggest that dialogue interlocutors reflect on their current speech, and that emotions in dialogue are not unidimensional. After modifying the model, a faster decrease in the loss function can be observed under identical conditions, leading to superior outcomes. Nevertheless, the revamped model also presents certain issues, such as an over-representation of neutral emotions, which results in erroneous classification of neutral emotions, and incorrect identification of similar emotions. These issues require remediation in future research.

References

[1] Hu, Dou, Lingwei Wei, and Xiaoyong Huai. "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations." arXiv preprint arXiv:2106.01978 (2021).

[2] Yang, Lin, et al. "Hybrid curriculum learning for emotion recognition in conversation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 10. 2022.

[3] Shen, Weizhou, et al. "Directed acyclic graph network for conversational emotion recognition." arXiv preprint arXiv:2105.12907 (2021).

[4] Majumder, Navonil, et al. "Dialoguernn: An attentive rnn for emotion detection in conversations." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[5] Lee, Joosung, and Wooin Lee. "CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation." arXiv preprint arXiv:2108.11626 (2021).

[6] Ghosal, Deepanway, et al. "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation." arXiv preprint arXiv:1908.11540 (2019).

[7] Zhang, Duzhen, Feilong Chen, and Xiuyi Chen. "Dualgats: Dual graph attention networks for emotion recognition in conversations." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.

[8] Jing, Shaoling, Xia Mao, and Lijiang Chen. "Automatic speech discrete labels to dimensional emotional values conversion method." IET Biometrics 8.2 (2019): 168-176.

[9] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42 (2008): 335-359.

Wenxue Zhang, Postgraduate student. Her current research interests include multimodal emotion recognition.

Baoshan Sun, Associate Professor Sun Baoshan, Doctor of Engineering, Master Supervisor, Deputy Head of Department. The country sent abroad to study in the UK visiting scholar, CCF Member of China Computer Society. Graduated from Tianjin University with a master's degree in computer application technology, and graduated from Tianjin Polytechnic University in computer testing Ph.D. graduate. Selected into the "Outstanding Young Teachers Funding Program" of Tianjin Universities. Leading the research team in related research fields and achieved a series of scientific research results in research topics. And has been in foreign SCI journals, EI journals, domestic important journals and international more than 20 high-level academic papers have been published at the conference.