ESCP: Enhancing Emotion Recognition in Conversation with Speech and Contextual Prefixes

Xiujuan Xu, Xiaoxiao Shi, Zhehuan Zhao, Yu Liu*
School of Software Technology, Dalian University of Technology
xjxu@dlut.edu.cn, xiaoxiao_shi@mail.dlut.edu.cn, {z.zhao, yuliu}@dlut.edu.cn

Abstract

Emotion Recognition in Conversation (ERC) aims to analyze the speaker's emotional state in a conversation. Fully mining the information in multimodal and historical utterances plays a crucial role in the performance of the model. However, recent works in ERC focus on historical utterances modeling and generally concatenate the multimodal features directly, which neglects mining deep multimodal information and brings redundancy at the same time. To address the shortcomings of existing models, we propose a novel model, termed ***E**nhancing Emotion Recognition in Conversation with **S**peech and **C**ontextual **P**refixes* (ESCP). ESCP employs a directed acyclic graph (DAG) to model historical utterances in a conversation and incorporates a contextual prefix containing the sentiment and semantics of historical utterances. By adding speech and contextual prefixes, the inter- and intra-modal emotion information is efficiently modeled using the prior knowledge of the large-scale pre-trained model. Experiments conducted on several public benchmarks demonstrate that the proposed approach achieves state-of-the-art (SOTA) performances. These results affirm the effectiveness of the novel ESCP model and underscore the significance of incorporating speech and contextual prefixes to guide the pre-trained model.

Keywords: Emotion Recognition in Conversation, Prefix-tuning, Multimodal

1. Introduction

Emotions are an important part of human social activities. Precise identification and understanding of participants' emotional states in conversations are the basis of emotion-aware and emotion-driven applications.

ERC has a difference relative to traditional multimodal sentiment analysis in that ERC requires the modeling of complex emotional dependencies (Poria et al., 2019b). The emotional dynamics in a conversation consist of two properties: self-dependence and interpersonal dependence (Morris and Keltner, 2000). Self-dependence mainly considers the speaker's influence on his/her own emotions, and interpersonal dependence mainly considers the influence of other participants on the speaker's emotions.

Most existing works (Majumder et al., 2019; Ghosal et al., 2019) tend to construct complex networks to model historical dialogue structures, and usually perform shallow concatenation of multimodal features to perform multimodal fusion without sufficient extraction of modal interaction features. Due to the heterogeneity among various modalities, the concatenation operation makes it difficult to fully exploit the complementary information of each modality, and it is more likely to introduce redundant information instead. How to effectively fuse multimodal features and learn meaningful representations has been a key issue in multimodal research (Lin and Hu, 2022; Zadeh et al., 2018), which illustrates the importance and complexity of multimodal fusion, so it is necessary to explore effective modeling of both kinds of information when considering ERC.

Therefore, we propose the ESCP model, which focuses on both the complementarity of different modal information and also models the impact of historical conversations on the target utterance. We assume that emotions in conversations are influenced by three major aspects: self-dependence, interpersonal dependence, and multimodal information.

Specifically, to capture the influence of self-dependence on emotion, we encode the speaker's historical utterances with a fixed window size using a pre-trained model. To capture the influence of interpersonal dependence, a directed acyclic graph (Schlichtkrull et al., 2018) is constructed based on the relationship of the utterances. Then, we obtain a contextual representation that aggregates the features of neighboring nodes. This representation is used as the contextual prefix of the pre-trained model (Li and Liang, 2021) to dynamically learn the semantic space features. For fusing multimodal features, the acoustic feature of the target utterance is used as the second prefix, which is responsible for providing complementary information of acoustic modality to the text in the process of fine-tuning the pre-trained model. These three parts of information are fused within the pre-training model through a multi-headed attention mechanism to obtain the final fused features for emotion prediction.

* Corresponding author

LREC-COLING 2024, pages 6278–6287

We conducted extensive experimental evaluations on two public benchmark datasets, IEMOCAP and MELD. The experimental results show that our model has significant performance advantages. Our contributions can be summarized in the following three aspects:

- Proposed a novel model Enhancing Emotion Recognition in Conversation with Speech and Contextual Prefixes that addresses the shortcomings of existing models, and to the best of our knowledge, we are the first work to use the prefix-tuning method in ERC.
- A prefix model for modeling historical conversations is proposed, which can deeply model the emotional information.
- Our model is easily extended to conversations with more than two people.

The remainder of the paper is organized as follows: Section 2 reviews related work and presents research results related to multimodal emotion recognition; Section 3 describes our proposed model in detail; Section 4 offers the experimental design; Section 5 analyzes the results; and Section 6 concludes.

2. Related work

2.1. Emotion Recognition in Conversation

With the development of social media, vast amounts of multimedia data have been generated. Since contextual conversation plays an essential role in ERC, early approaches focus on historical conversation modeling, and specific approaches can be classified into two types based on Recurrent Neural Network (RNN) and Graph Neural Network (GNN).

RNN-based Models (Hazarika et al., 2018b) capture information that is meaningful for emotion recognition by encoding historical conversations as memory vectors and using attentional mechanisms to focus on important contextual segments. Since (Hazarika et al., 2018b) only considers the individual information of speakers and ignores the influence of other speakers on the target utterance, (Hazarika et al., 2018a) proposes to encode the historical conversations of all participants so that the model includes the interaction information of participants. (Majumder et al., 2019) uses three Gate Recurrent Unit (GRU) modules to separately perform three aspects of speaker information, historical conversation, and emotion of historical. Other RNN-based methods include (Bansal et al., 2022).

GNN-based Models Recent studies have found that utterances in dialogues can construct a directed graph based on information such as the chronological order of speech, different speakers, etc., and can converge graph node information through graph neural networks to model historical conversations and multimodal information. (Ghosal et al., 2019) solves the context propagation problem in the RNN-based approach through GNN, but the model does not focus on multimodal fusion and simply encodes speech, text, and visual modal information uniformly with GRU. (Hu et al., 2021b) considers multimodal features as nodes as well. (Shen et al., 2021) uses DAG to model the intrinsic structure of conversations, considering different speakers and utterance sequences. Inspired by this model, we designed the context encoder. Other GNN-based methods include (Lee and Choi, 2021; Hu et al., 2022).

2.2. Prefix-tuning

With the widespread success of pre-trained models, a number of techniques have been developed to tune pre-trained models with the expectation of transferring the generalization capabilities of large models to other downstream tasks. (Li and Liang, 2021) proposed "prefix-tuning" in 2021, which aims to guide models to generate text in a particular direction by inserting a specific prefix before the generation task. The core idea of "prefix-tuning" is to use a specialized prefix model to generate appropriate prefixes. (Arjmand et al., 2021) has achieved excellent performance on multimodal sentiment analysis tasks by using the output of a speech model as a prefix.

3. Methodology

The overall framework is shown in Figure 1. The concatenated utterances $\widetilde{u}$ are tokenized using a tokenizer, and the constructed directed acyclic graph is encoded with a context encoder to obtain the contextual prefix. The embedding with contextual prefix and speech prefix is used as the input to the pre-trained model. This section describes our approach in detail.

3.1. Problem Definition

A conversation contains a series of utterances that can be defined as $\{u_1, u_2, ..., u_N\}$, u_i denotes an utterance, and N denotes the number of utterances

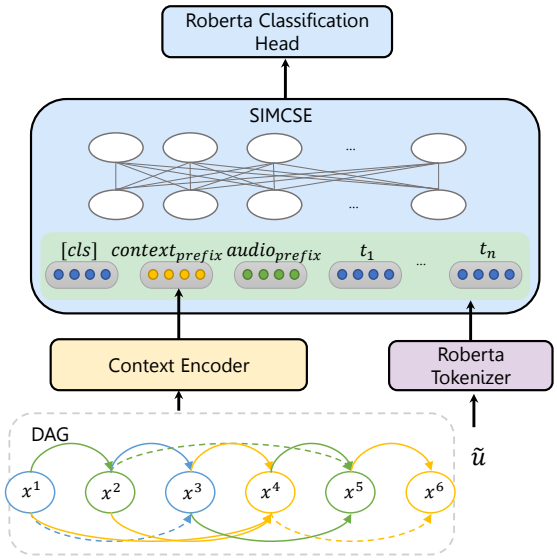


Figure 1: The architecture of ESCP model. Train the pre-trained model for classification using two prefixes. The window size for this example is 1. x is the textual feature of the utterance. Circles of the same color represent the same speaker. Solid and dotted lines represent two relationship types.

in the conversation. The set of speakers is defined as $\{s_1, s_2, ..., s_J\}$, indicating that there are J speakers, u_{i,s_j} refers to the ith utterance, and the speaker is s_j. $y_i \in Y$ is the emotion label of the utterance u_i, and Y is the set of emotion categories. The goal of the task is to predict the emotion class of the target utterance based on a set of utterances and speakers.

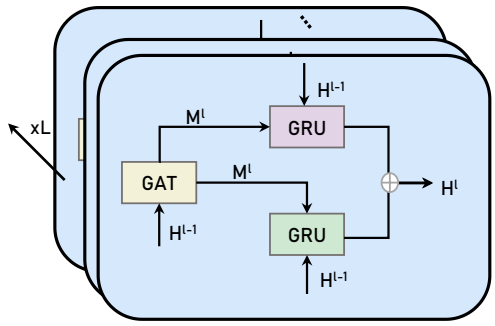


Figure 2: Framework illustration of the Context Encoder. Context Encoder contains x layers, H^{L-1} is the output of the previous layer, uses two GRUs to encode the output of GAT and H^{L-1}, and finally adds the output of the two GRUs to get the output of this layer.

3.2. Context Encoder

DAG-ERC (Shen et al., 2021) improves the aggregation function of Directed acyclic graph models (DAGNN) (Thost and Chen, 2021) by borrowing from Relation Graph Convolution Neural Networks (RGCN) (Schlichtkrull et al., 2018), considering the type of edges, using two GRU units to aggregate node features, and improving GAT (Velickovic et al., 2018). Our context encoder mainly uses this GAT variant to aggregate node features, obtaining the context encoding representation $context_{prefix}$, as the prefix of the pre-trained model. The structure of the context encoder is shown in Figure 2. Since text contains the main emotional information, we use the text features of utterances $\{x_1^t, x_2^t, ..., x_N^t\}$ for contextual encoding. Text features are extracted by BERT.

We represent each conversation as a directed acyclic graph $G = (V, E, R)$. V is the set of nodes, E is the set of edges, and R is the set of relationship types of edges. The nodes in the graph are represented by the textual features of each utterance, and the edges are divided into two types, noted as $\{0, 1\}$, where 1 means that the two nodes corresponding to this edge are the same speaker and 0 means a different speaker. The direction of the edges points from past utterances to future utterances, representing that only past information is used when predicting the target utterances, which is in line with the practical meaning. The number of edges can be controlled by the window size ω. The meaning of ω is to intercept a window from the location of the target utterance forward so that it contains at most ω utterances of the same speaker. Both the utterances within the window and the target utterance form an edge that establishes a link between the target utterance and the historical utterance.

The context encoder contains multiple layers, where features are updated by iteration. At each layer, the aggregation function of the GAT variant applies a relation-aware feature transformation to leverage the relationship types of edges while collecting information based on attention weights. At each layer, the hidden state of the utterance is computed cyclically over the time stream from the first utterance to the last utterance.

$$M_i^l = \sum_{j \in N_i} \alpha_{ij} W_r^l H_j^l \quad (1)$$

where W_r^l is the trainable parameter of the relation-aware transformation and H_j^l is initially the node feature. α_{ij} is the attention weight, which is calculated by the hidden state of the previous layer and the hidden state of the neighboring nodes in the current layer :

$$\alpha_{ij}^l = Softmax_{j \in N_i}(W_\alpha^l [H_j^l || H_i^{l-1}]) \quad (2)$$

where || is the concat operation.
After obtaining aggregated information M, the node information unit GRU_H and the contextual information unit GRU_M are made to interact with the utterance u_i.

$$\tilde{H}_i^l = GRU_H^l(H_i^{l-1}, M_i^l) \quad (3)$$

$$C_i^l = GRU_M^l(M_i^l, H_i^{l-1}) \quad (4)$$

The final representation of u_i in the l-layer is the sum of $\tilde{H}_i^l$ and C_i^l:

$$H_i^l = \tilde{H}_i^l + C_i^l \quad (5)$$

Finally, the hidden state of each layer is concatenated to obtain $context_{prefix}$.

$$context_{prefix} = ||_{l-0}^{L} H_i^l \quad (6)$$

3.3. Prefix-tuning

The output obtained from the text features after the context encoder is noted as $context_{prefix}$. Considering that the emotion of the target utterance is influenced by self-dependence, we concatenate the first K utterances of the same speaker to reinforce the emotional information contained in the utterance. It has been demonstrated that nonverbal modality can provide complementary information to textual modality, so we concatenate the acoustic features of the target utterance as prefixes into the input of SIMCSE[1] (Gao et al., 2021) as well. The audio features are extracted using the tool OpenSMILE. 6373 features were extracted for each utterance and then reduced to 100 dimensions using a fully connected layer. The concatenated text is $\tilde{u} = \{u_{0,s_j}, ..., u_{k-2,s_j}, u_{k-1,s_j}, u_{i,s_j}\}$. After $\tilde{u}$ has been represented as $\{t_1, t_2, ...t_n\}$ by the embedding layer of the pre-trained model, the contextual prefix $context_{prefix}$ and the acoustic prefix $audio_{prefix}$ are concatenated behind the CLS token to obtain $e = \{cls, context_{prefix}, audio_{prefix}, t_1, t_2, ...t_n\}$. e is used as input to SIMCSE to fine-tune the model and predict the emotion class using the hidden state of the last layer. For the training of the model, we use the standard cross-entropy loss as the objective function:

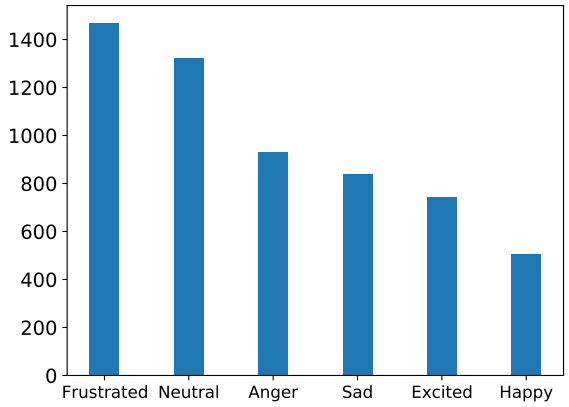
$$L(\theta) = -\frac{1}{N}\sum_{i}^{N}\sum_{c=1}^{M} y_{ic} log(p_{ic}) \quad (7)$$

where N is the number of training conversations, M is the number of utterances in the i_{th} conversation, y_{ic} is the truth label, p_{ic} is the probability , and θ is the set of trainable parameters of the model.

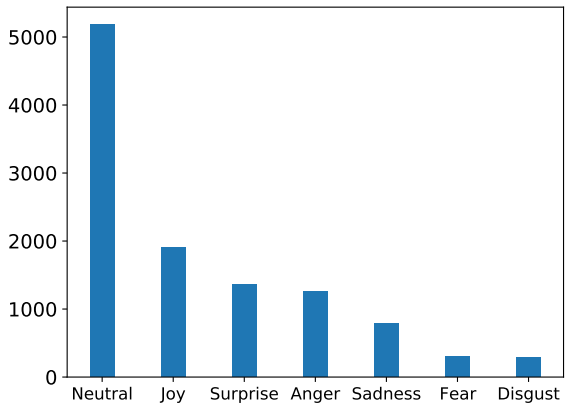
[1]SIMCSE is a model for learning sentence embeddings. The core idea of the SIMCSE model is to use a contrast loss function to train sentence embeddings, which can map similar sentence embeddings to similar locations to form sentence representations with semantic information. Since the pre-trained model has the natural ability to model context, we use the pre-trained SIMCSE model as a backbone network to fuse speech prefixes, contextual prefixes, and speaker-level historical utterances to obtain the final fused feature representation. Reference:sup-simcse-roberta-base

4. Experimental Setups

This section details the datasets used in the experiment, the compared models, and the settings of the hyperparameters.



(a) IEMOCAP



(b) MELD

Figure 3: Statistics on the number of samples in each category of the IEMOCAP and MELD dataset.

4.1. Datasets

IEMOCAP (Busso et al., 2008) IEMOCAP is a widely used multimodal dataset collected by the

IEMOCAP								
Model	**Happy**	**Sad**	**Neutral**	**Angry**	**Excited**	**Frustrated**	**Accuracy**	**wa-F1**
DialogueRNN(Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN(Ghosal et al., 2019)	42.75	84.54	63.54	64.19	63.08	66.99	65.25	64.18
DAG-ERC(Shen et al., 2021)	70.21	62.64	67.35	76.11	50.60	53.42	65.04	65.15
DialogueCRN(Hu et al., 2021a)	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
MMGCN(Hu et al., 2021b)	42.34	78.67	61.73	69.00	74.33	62.32	-	66.22
GraphCFC(Li et al., 2022a)	43.08	84.99	64.70	71.35	78.86	63.70	69.13	68.91
SPCL-CL-ERC(Song et al., 2022)	69.53	63.79	60.97	**80.34**	47.06	67.66	66.46	66.59
EmoCaps(Li et al., 2022b)	**70.41**	82.72	64.38	65.31	78.75	65.64	70.65	70.68
ESCP (ours)	70.00	**85.90**	**79.79**	75.82	**83.25**	**73.77**	**78.77**	**78.69**

Table 1: Results on the IEMOCAP dataset. Bolded font means the best results. Wa-F1 means weighted average F1 score and is also the final evaluation metric. The result of DAG-ERC is the best result obtained by running it five times. The results of other models come from the original paper or the results in other papers.

MELD							
Model	**Neutral**	**Surprise**	**Sadness**	**Joy**	**Anger**	**Accuracy**	**wa-F1**
DialogueRNN(Majumder et al., 2019)	76.79	47.69	20.41	50.92	45.52	60.31	57.66
DialogueGCN(Ghosal et al., 2019)	-	-	-	-	-	-	58.10
DAG-ERC(Shen et al., 2021)	76.15	**54.23***	23.69	**57.08***	46.86	**63.62***	62.17
DialogueCRN(Hu et al., 2021a)	76.13	46.55	11.43	49.47	44.92	59.66	56.76
MMGCN(Hu et al., 2021b)	-	-	-	-	-	-	58.65
GraphCFC(Li et al., 2022a)	**76.98***	49.36	26.89	51.88	47.59	61.42	58.86
SPCL-CL-ERC(Song et al., 2022)	**80.16**	**57.45**	**43.45**	**64.48**	**51.74**	**68.53**	**68.07**
EmoCaps(Li et al., 2022b)	75.01	48.24	26.11	51.23	42.97	60.28	59.64
ESCP (ours)	76.64	50.50	**29.13***	55.64	**48.14***	62.77	**62.39***

Table 2: Results on the MELD dataset. DialogueGCN and MMGCN do not have detailed wa-F1 for each emotion category. Bolded font means the best results. * refer to our results are the best compared with other models except for SPCL-CL-ERC. The results of DAG-ERC, SPCL-CL-ERC, and EmoCaps are the best results obtained by running five times. The results of other models come from the original paper or the results in other papers.

Dataset	Partition	No.Uttrs	No.Dials
IEMOCAP	train + val	5810	120
	test	1564	31
MELD	train + val	11098	1152
	test	2610	280

Table 3: Statistics of the two datasets.

University of Southern California, containing audio, transcriptions, video, and motion-capture(MoCap). It contains more than 150 conversations in 10 emotion categories.

MELD (Poria et al., 2019a) MELD is a clip taken from the TV series Old Friends. It is a multimodal dataset that includes both text, audio, and video information. MELD has over 1400 dialogue pairs with a total of 13,000 utterances. It contains 7 emotions, namely Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear.

Table 3 lists the statistics of the dataset. No.Uttrs represents the number of utterances, No.Dials represents the number of dialogues.

4.2. Implementation Details

The hyperparameters are set as follows: on both IEMOCAP and MELD datasets, the context encoder is 4 layers, the hidden layer size is 300, due to the unbalanced data category of MELD, the focal loss is used on MELD, the gamma is set to 2. The learning rate of the context encoder is 1e-4, and the learning rate of the SIMCSE model is 1e-6. IEMOCAP was trained for 100 epochs and MELD was trained for 50 epochs. The machine used is an NVIDIA GeForce RTX 3090 Ti. We use five seeds and report the best result.

4.3. Baselines and State of the Art

We compare our model with baselines, described as follows:

DialogueRNN (Majumder et al., 2019) is a renowned sequence-based model that uses three GRUs to model the speaker, context, and the previous sentiment. **DialogueGCN (Ghosal et al., 2019)** uses fusion features as nodes that establish dependencies between itself and other speakers. **MMGCN (Hu et al., 2021b)** treats multimodal information as nodes of the graph as well. **DAG-ERC**

(Shen et al., 2021) modeling dialogues using directed acyclic graphs combines the advantages of GNN and RNN. **DialogueCRN (Hu et al., 2021a)** learning of LSTM features using a multi-round attention mechanism. **GraphCFC (Li et al., 2022a)** fully Integrates features for each pair of modals. **SPCL-CL-ERC (Song et al., 2022)** State-of-the-art on the MELD dataset using the prototype contrast learning to solve the imbalance classification problem. **EmoCaps (Li et al., 2022b)** leads in performance on the IEMOCAP dataset, employing a transformer encoder structure for feature extraction in each modality.

4.4. Evaluation Metrics

Due to the large variation in the number of samples from different emotion categories in each dataset, using accuracy as a metric is not informative, so we use the weighted average F1-score as the final evaluation metric. Weighted F1-score is the weighted summed average of Precision and Recall, which is a common evaluation criterion in the field of information retrieval (IR) and can fairly evaluate the classification model's overall performance.

5. Results and Discussions

This section introduces experimental results on two datasets and ablation experiments for different modules. Then discuss the impact of context window on the model. And the effect of the model is intuitively reflected through visualization. Finally, a detailed analysis of the erroneous samples is performed.

5.1. Comparison with State of the Art and Baseline

IEMOCAP Table 1 shows the experimental results on the IEMOCAP dataset. On the IEMOCAP dataset, our model outperforms the SOTA model by 11.49% in Accuracy and 11.34% in Weight F1 and achieves the best performance on Sad, Neutral, Excited, Frustrated, and competitive performance on Happy, and Angry. This shows that our model is effective.

MELD Table 2 shows the experimental results on the MELD dataset, and SPCL-CL-ERC is SOTA on this dataset, this model aims to solve the problem of category imbalance in the dataset, and it is very effective. The distribution of the number of categories in IEMOCAP and MELD is shown in Figure 3, and it can be seen that the number of categories in MELD is very different. The performance of our model does not outperform SOTA on MELD, which indicates that ESCP is less capable of dealing with the category imbalance dataset, nevertheless, compared with other models, ESCP still has competitive performance.

5.2. Effect of Context Window

We concatenated K utterances of the same speaker for the target utterance to enhance the emotional information contained in the utterance, and to investigate the effect of the number of concatenated utterances, we conducted experiments with different K values. The results are shown in Table 4. The results show that the performance of concatenating 1,2,3 utterances is similar. But when concatenating four utterances the performance of the model decreases instead. This may be because the historical utterances that are too far away do not have much influence on the target utterances and the sentiment may have shifted. When only the target utterance is used, the performance is degraded because the utterance is too short.

5.3. Ablation Study

We conducted ablation experiments on prefixes, and the results are displayed in Table 5. We observe a minor decrease in performance, with the f1 score decreasing by 0.94%, when removing the speech prefix. This indicates that the speech prefix contributes multimodal and complementary information. However, the performance is significantly reduced when the contextual prefix is removed, with the f1 score decreasing by 34.77%. After removing all prefixes is equivalent to directly fine-tuning SIMCSE, the performance is the worst, with the f1 score down by 36.66%. The results illustrate the very large contribution of context to the model. Combined with previous work, we believe this is because, in ERC, the sentiment of the target utterance is influenced by the emotion of the historical conversation.

5.4. Visualization

To study the effect of prefixes on features, we visualized the hidden state of the last layer of the SIMCSE model by downscaling. Both plots were produced using the t-SNE algorithm with 20000 iterations and perplexity set to 30. Figure 4 is the result without prefixes, the Figure 5 is the result of the ESCP model, and we can see that the semantic features extracted with prefix are more closely located between the same category, the category boundary is more clear, the features without prefix are more scattered, multiple categories are mixed, it is difficult to distinguish. In ESCP we observed that happy and excited emotions are more similar, so they are spatially closer to each other, while negative emotions such as sadness, anger, and frustration are farther away. Experimental data also

K	Happy	Sad	Neutral	Angry	Excited	Frustrated	Accuracy	wa-F1
0	**74.49**	84.78	78.26	70.71	83.13	72.75	77.35	77.74
1	71.88	84.93	79.21	73.58	83.70	**74.90**	78.71	78.67
2	70.00	**85.90**	79.79	**75.82**	83.25	73.77	**78.77**	**78.69**
3	72.44	82.57	**81.27**	73.65	**83.85**	73.32	78.58	78.55
4	69.88	84.08	77.91	72.05	81.90	73.91	77.43	77.33

Table 4: Effect of concatenating different numbers of utterances on IEMOCAP dataset.

Model	Happy	Sad	Neutral	Angry	Excited	Frustrated	Accuracy	wa-F1
ours	**70.00**	85.90	**79.79**	**75.82**	**83.25**	73.77	**78.77**	**78.69**
w/o Audio_prefix	69.54	**86.27**	78.40	73.14	82.11	**74.10**	78.01	77.95
w/o Context_prefix	5.23	55.36	58.09	48.92	57.90	54.57	53.26	51.30
w/o all prefix	9.27	53.76	54.62	46.02	60.07	50.83	51.53	49.81

Table 5: Results of the ablation experiments on the IEMOCAP dataset.

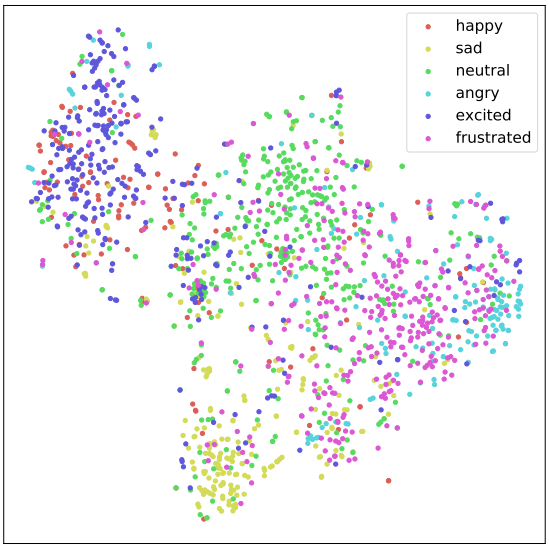


Figure 4: Reduced dimensional visualization of SIMCSE hidden states without prefixes, the boundaries are very fuzzy, and the categories are crossed together.

Figure 5: Reduced dimensional visualization of SIMCSE hidden states with prefixes, clear boundaries, samples of the same category clustered together, positive emotions, and negative emotions farther apart.

show that similar emotions are more likely to be misclassified.

To investigate the contribution of prefixes to the model, we visualized the attention weights of the last layer of the SIMCSE model shown in Figure 6 and Figure 7. Both examples of models recognized correctly. We found that the model gives high attention to both contextual prefixes and acoustic prefixes, which indicates that prefixes provide a large amount of effective information. Speech information also provides a more important contribution than context when the contextual prefix emotion information is not obvious. For example, sometimes in angry or excited emotions, the acoustic features of people, are different from normal speech. In Figure 6, although the model also noticed the negative word "cried", it paid more attention to the prefixes, especially the speech prefixes, and finally correctly classified the utterance as happy.

5.5. Error Analysis

To explore the effects of emotion shifts, we counted the samples of misclassified in the test set. Emotion shifts are the change of emotion from one category to another within a conversation. The test set contains 30 dialogues with a total of 1564 utterances. Among them, 350 utterances are incorrectly classi-

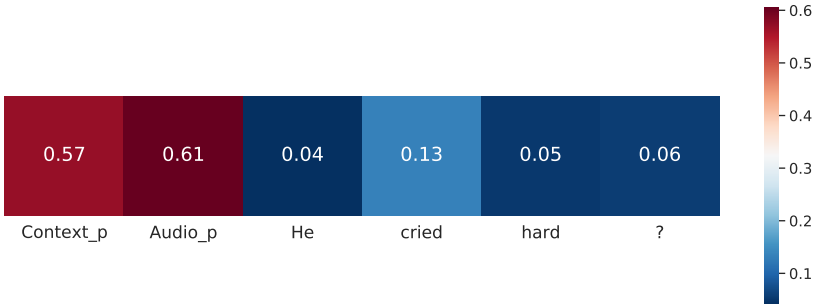


Figure 6: Visualization of the attention weight corresponding to utterance with happy emotion.

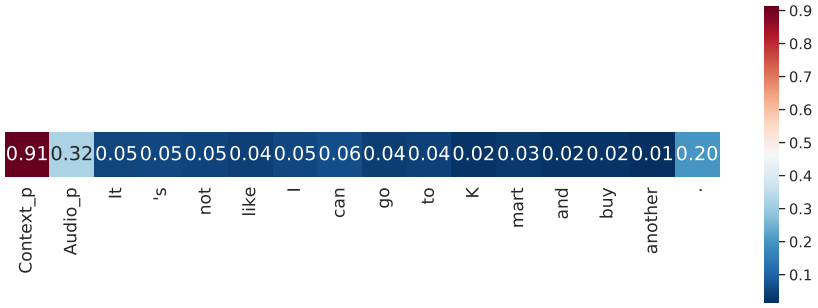


Figure 7: Visualization of the attention weight corresponding to utterance with angry emotion.

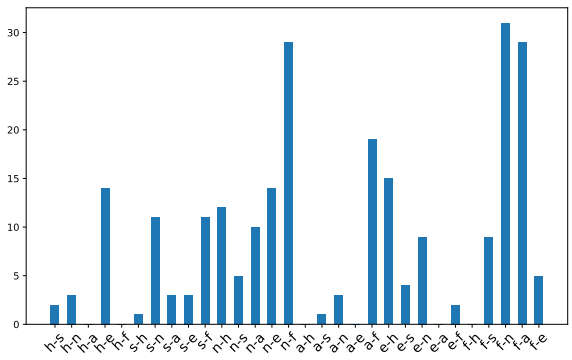


Figure 8: Number of samples of recognition errors caused by 30 emotion shifts in IEMOCAP

fied. Among these 350 error utterances, 245 are due to emotion shifts. The statistics found that the number of samples misclassified due to emotion shifts is 70%, which indicates that the identification of emotion shifts may be a difficulty for the model. Figure 8 shows more detailed statistics of emotion shifts. The vertical axis is the number of utterances with incorrect recognition. The letters on the horizontal axis are the first letters of the emotion categories: $\{happy : h, sad : s, neutral : n, angry : a, excited : e, frustrated : f\}$. $a - b$ is the number of samples where the emotion of the previous utterance is a, the emotion of the target utterance is b, and the target utterance is incorrectly predicted. The larger the number of samples, the more difficult it is to identify the transfer between these two emotions. From the figure 8, we can see that $neutral$ to $frustrated$, $frustrated$ to $neutral$, and $frustrated$ to $angry$ all have more samples, which may be due to the weak distinction between the emotions of frustrated and neutral.

6. Conclusion

To balance multimodal fusion and historical dialogue modeling, we use prefix-tuning to fine-tune the pre-trained model and migrate the capabilities of the large model to our proposed model, addressing the shortcomings of the existing ERC model. We also propose a prefix model for modeling historical dialogues that can dynamically learn semantic space features. We evaluate our proposed ESCP on two benchmark datasets widely used by most ERC models. Experimental results show that our proposed model can effectively perform intra- and inter-modal interactions to extract contextual and complementary information. The ESCP has sig-

nificant advantages over previous baseline models. In the future, we will explore solving the problem of poor performance of the model on category-imbalanced datasets.

7. Acknowledgements

This work is funded in part by the National Natural Science Foundation of China Project(No.62372078, 62376049) and the Dalian Key Field Innovation Team Support Plan (Grant: 2020RT07).

8. Bibliographical References

Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. TEASEL: A transformer-based speech-prefixed language model. *CoRR*, abs/2109.05522.

Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56, Virtual. International Conference on Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. 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*, pages 7037–7041. IEEE.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. 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 (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online. Association for Computational Linguistics.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2022a. Graphcfc: A directed graph based cross-modal feature complementation approach

for multimodal conversational emotion recognition. *CoRR*, abs/2207.12261.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022b. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Ronghao Lin and Haifeng Hu. 2022. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 511–523, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Michael W Morris and Dacher Keltner. 2000. How emotions work: The social functions of emotional expression in negotiations. *Research in organizational behavior*, 22:1–50.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Veronika Thost and Jie Chen. 2021. Directed acyclic graph neural networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5634–5641. AAAI Press.