Contrast and Generation Make BART a Good Dialogue Emotion Recognizer

Shimin Li[1,3], **Hang Yan**[1,3], **Xipeng Qiu**[1,2,3*]

[1] School of Computer Science, Fudan University
[2] Peng Cheng Laboratory, Shenzhen, Guangdong, China
[3] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{smli20, hyan19, xpqiu}@fudan.edu.cn

Abstract

In dialogue systems, utterances with similar semantics may have distinctive emotions under different contexts. Therefore, modeling long-range contextual emotional relationships with speaker dependency plays a crucial part in dialogue emotion recognition. Meanwhile, distinguishing the different emotion categories is non-trivial since they usually have semantically similar sentiments. To this end, we adopt supervised contrastive learning to make different emotions mutually exclusive to identify similar emotions better. Meanwhile, we utilize an auxiliary response generation task to enhance the model's ability of handling context information, thereby forcing the model to recognize emotions with similar semantics in diverse contexts. To achieve these objectives, we use the pre-trained encoder-decoder model BART as our backbone model since it is very suitable for both understanding and generation tasks. The experiments on four datasets demonstrate that our proposed model obtains significantly more favorable results than the state-of-the-art model in dialogue emotion recognition. The ablation study further demonstrates the effectiveness of supervised contrastive loss and generative loss.

Introduction

With the development and popularization of personal intelligent terminal technology and social networks, the importance of constructing a dialogue system that can comprehend user emotions and intentions and conduct effective dialogue interactions has increased significantly. A critical module in the dialogue system is the natural language understanding module that analyzes user behaviors like intents or emotions. Analyzing user sentiments with contextual relationships is an advanced step for simple sentiment classification tasks and is more suitable for usage scenarios in the real world with more research value. The task of emotion recognition in conversation (ERC) is to assign emotion labels to all the utterances in a historical dialogue with a contextual relationship. At the same time, each historical dialogue contains interactions between multiple different speakers, as illustrated in Figure 1.

There are three challenges for ERC. (1) The first challenge is that the emotion of each utterance may be affected by contextual information. For example, specific emotions will depend on certain utterances of the context. Meanwhile, utterances with the same expression may have completely different emotions in various contexts. Therefore, effectively modeling the context dependency and the speaker dependency is the main factor distinguishing this task from traditional sentiment classification. (2) The second challenge is that each speaker's emotion is influenced by the utterance of other speakers in the conversation, so there may exist a sudden change in a speaker's emotion. (3) The third challenge lies in semantically similar but different categories of emotions, such as "frustrated" to "sad", "happy" to "excited", etc. It is difficult to distinguish these semantically similar sentiment categories.

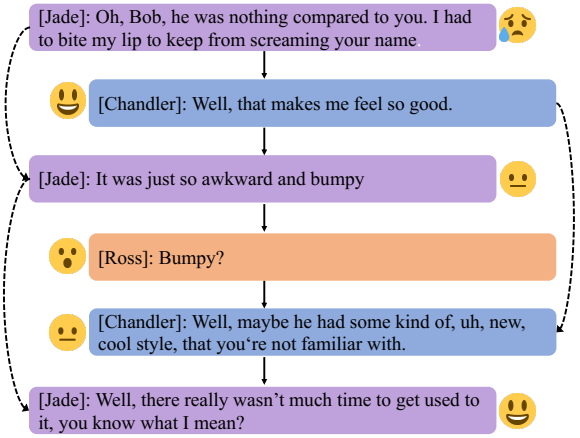

Figure 1: The conversation flow chart in multi-person dialogue emotion recognition. The solid line indicates that the previous utterance directly influences the current speaker's emotion. The dashed line signifies that the same speaker is influenced by other utterances and expresses different emotions.

Recent related work addressed contextual dependencies and speaker relations using various graph networks (Shen et al. 2021b; Ghosal et al. 2019; Ishiwatari et al. 2020; Sheng et al. 2020). However, as the number of layers deepens, the phenomenon of over-smoothing (Chen et al. 2020a) starts to appear, resulting in the representation of similar sentiments

*Corresponding Author.

tending to be indistinguishable.

This work deals with the above challenges by better modeling the context and speaker information and auxiliary generation task.

Firstly, we introduce a dialogue-level Transformer (Vaswani et al. 2017) layer to model the long-range context dependencies between utterances. A pre-trained language model captures the representation of each utterance. Compared to previous approaches that only adopt pre-trained models as a feature extractor (Liu et al. 2019) and employ the extracted features as the node representation of downstream graph networks, a pure Transformer structure makes fewer prior structural assumptions (Lin et al. 2021).

Secondly, we adopt supervised contrastive learning (SCL) (Khosla et al. 2020) to alleviate the difficulty in categorizing similar emotions, which makes samples with same sentiments cohesive and different sentiments mutually exclusive under the fully utilization of label information. Compared with the cross-entropy loss for noisy labels, the supervised contrastive loss can increase the stability of training and improve the generalization of the model (Gunel et al. 2021). Unlike the regular SCL, we copy the hidden state of all samples in a batch and detach off its gradient as its multiview representation. The reason is that the categories in existing ERC datasets are highly unbalanced, and some categories may exist in a batch with only one sample. If only the original SCL is used, it will lead to incorrect loss calculation.

Thirdly, we introduce an auxiliary response generation task to enhance the ability of capturing the context information for ERC. The prediction of the following utterance makes the model fully consider contextual dependencies, thus forcing the model to consider the information in the context and rely on the current utterance itself when recognizing the sentiment in the conversation. Moreover, by splicing the speaker directly before utterance as a hint for speaker information, the dependency between speakers and utterances is modeled adequately without additional parameters.

Finally, we utilize BART (Lewis et al. 2020), a pre-trained Transformer with an encoder-decoder structure, as our backbone model and enhance it by contrastive and generative loss. Our proposed **CO**nstrastive-and-**G**eneration-enhanced BART (CoG-BART) obtains state-of-the-art results on four ERC datasets compared to the baseline models. Additionally, ablation experiments and case studies prove the effectiveness of the contrastive and generative losses in the ERC task[1].

To summarize, our main contributions can be concluded as follows:

- To the best of our knowledge, we utilize supervised contrastive learning for the first time in ERC and significantly improve the model's ability to distinguish different sentiments.
- By incorporating response generation as an auxiliary task, the performance of ERC is improved when certain contextual information is involved.
- Our model is easy-to-implemented since it does not depend on external resources, like graph-based methods.

[1]https://github.com/whatissimondoing/CoG-BART.

Related Work

This section will introduce related works in ERC. Due to context-dependency and speaker dependency properties, it is natural for researchers to employ graph neural networks. Therefore, many works have constructed various task-specific graph structures. Meanwhile, with the excellent performance of the pre-trained model in diverse downstream tasks, an increasing number of works adopt the pre-trained model as the feature extractor for the input of the downstream model or directly fine-tune it with downstream datasets. Therefore, this section divides the related work into two categories: graph-based models and pre-train-based models.

Dialog Emotion Recognition

Graph-based model Considering the unidirectionality of information interaction, DAG (Shen et al. 2021b) utilizes directed acyclic graphs to model the information interaction between utterance and speaker. DialogGCN (Ghosal et al. 2019) adopts the basic graph neural network to model the relationship between contexts. SumAggGIN (Sheng et al. 2020) adds an aggregation module based on DialogGCN to additionally consider phrase-level information other than utterance-level. By simulating the process of human reasoning, DialogCRN (Hu, Wei, and Huai 2021) proposes to apply several reasoning modules to extract and integrate clues of emotional reasoning. To make the model better understand the additional general information involved in the dialogue process, KET (Zhong, Wang, and Miao 2019) combines external knowledge with a hierarchical Transformer. By appending sequence information into the graph network, RGAT (Ishiwatari et al. 2020) uses relational position encoding to combine position information into the graph network structure to consider the dependency between speakers. TODKAT (Zhu et al. 2021) integrates topic detection into the pre-training model and fuses commonsense knowledge into Transformer (Vaswani et al. 2017).

Pre-train-based model Suppose each utterance is regarded as an independent sentence, regardless of its context-dependence and speaker information. In that case, the problem can be transformed into a simple sentence classification so that pre-trained models (Qiu et al. 2020) such as BERT (Devlin et al. 2019), BART (Lewis et al. 2020), and RoBERTa (Liu et al. 2019) can be used directly for fine-tuning. HiTrans (Li et al. 2020) adopts BERT to extract utterance features, followed by transformer structure for modeling context. Considering speaker dependence, the auxiliary task of judging whether two utterances are the same speaker is used to model the speaker information. COSMIC (Ghosal et al. 2020) exploits RoBERTa as the feature extractor of each utterance and model the dependency of the context with RNN. In addition, the common knowledge transformer COMET (Bosselut et al. 2019) is incorporated to introduce world knowledge. Based on XLNet, DialogXL (Shen et al. 2021a) changes the segment-level

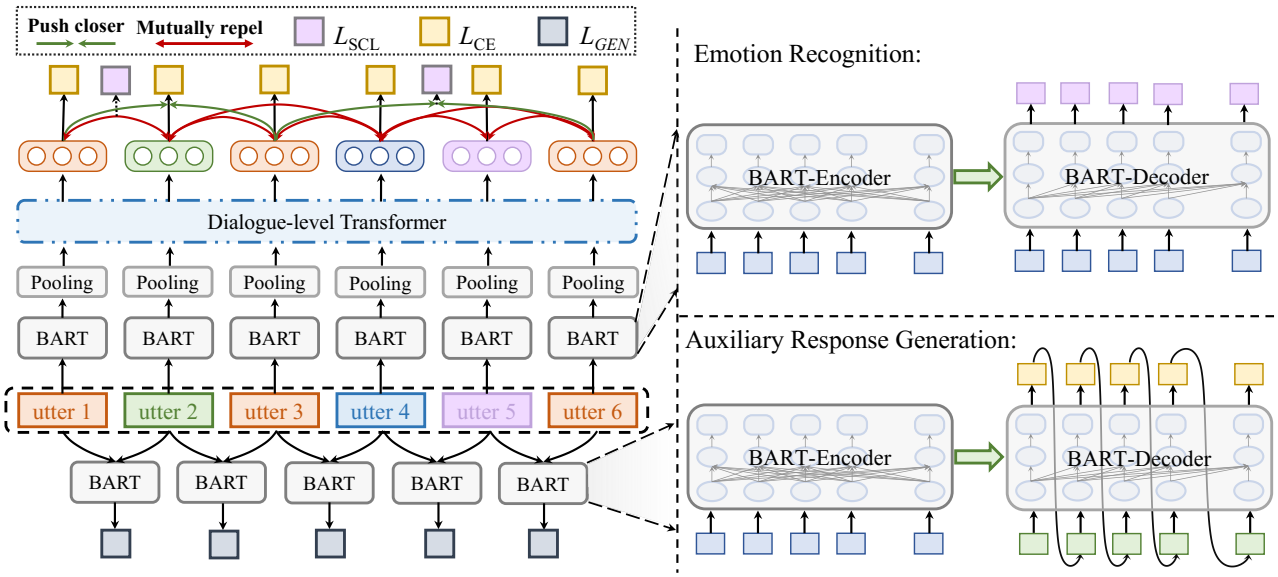

Figure 2: The overall framework of CoG-BART. The utterance is fed into BART for N utterances in a batch to get its hidden state. The representation of the utterance obtained after max-pooling the hidden state of each utterance is fed to the upper-level dialogue-level Transformer for modeling context dependencies. The obtained context-dependent utterance representations are utilized to compute the cross-entropy loss and supervised contrastive loss. In addition, the two adjacent utterance pairs are used for the auxiliary response generation.

structure to utterance-level and uses memory to record the historical context. Meanwhile, by adopting different mask mechanisms on different attention heads, each attention head pays attention to different aspects of dialogue information. Ide and Kawahara (2021) trained BART with both generation and classification in a multi-task format, though their method focused on response generation, treating emotion recognition as an auxiliary task. However, we focus on ERC and apply supervised contrastive loss as an additional optimization goal.

Contrastive Learning

Unsupervised contrastive learning In the field of computer vision, SimCLR (Chen et al. 2020b) takes pictures obtained from the same image through randomly different data augmentation methods as positive samples and other pictures as negative samples, thereby optimizing contrastive loss. The naive sentence representation obtained by BERT has poor performance in semantic text similarity tasks. Therefore, ConSERT (Yan et al. 2021) introduces self-supervised contrast loss in the fine-tuning stage of BERT. MBERT (Kim, Yoo, and Lee 2021) does not use data augmentation to construct positive samples but uses BERT with frozen parameters and fine-tunable parameters as a special siamese model to construct positive samples.

Supervised contrastive learning To make full use of label information, Khosla et al. (2020) extends it to supervised contrastive learning based on self-supervised training so that samples belonging to the same label are gathered in the embedding space while pushing samples of different categories away. Given that cross-entropy loss may cause model training instability and converge to a local optimum, SCL (Gunel et al. 2021) introduces supervised contrastive loss in the fine-tuning stage, which greatly improves the model's performance in few-shot learning scenarios. SimCSE (Gao, Yao, and Chen 2021) uses entailment pair in the annotated NLI dataset as the positive sample and the contradict pair as the negative sample in supervised contrastive learning.

Methodology

Problem Definition

In dialogue emotion recognition, the data is composed of multiple conversations $\{c_1, c_2, \cdots, c_N\}$, with each conversation composed of several utterances $c_i = [u_1, u_2, \cdots, u_m]$ and emotion labels $\mathcal{Y}_{c_i} = \{y_1, y_2, \cdots, y_m\} \in S$, where S indicates the categories of emotions. For an utterance, it is comprised of several tokens $u_t = [w_{t,1}, w_{t,2}, \cdots, w_{t,n}]$. Every utterance in a conversation c_i is uttered by one speaker which can be represented as $p(c_i) = [p(u_1), \cdots, p(u_i), \cdots, p(u_m)]$ and $p(u_i) \in P$, where P indicates the categories or names of the speakers. Accordingly, the whole problem can be expressed as getting the emotional label of each utterance based on the context and speaker information in a piece of conversation: $\mathcal{Y}_{c_i} = f(c_i, p(c_i))$.

Supervised Contrastive Learning for ERC

Utterance Encoding To model the dependencies between speaker and utterance, for a certain utterance u_t in a conversation, we splice the speaker's name or category before the utterance. After tokenizing the utterance prepended with the speaker information, we get:

$$\tilde{u}_t = \left[\langle s\rangle, w_{t,1}, \cdots, w_{t,i}, \cdots, w_{t,|n_t|}, \langle /s\rangle\right], \quad (1)$$

where $\langle s\rangle$ and $\langle /s\rangle$ are treated as special tokens to indicate the beginning and end of an utterance. Then the token sequence after tokenization is fed to the shared embedding layer of BART to acquire the hidden state of each token in utterance before sending it to the encoder and decoder of BART. After sending H_t to BART, the representation of the current utterance $\widehat{H}_t$ is acquired:

$$H_t = \text{EmbeddingLayer}(\tilde{u}_t), \tag{2}$$

$$\widehat{H}_t = \text{BART-Model}(H_t), \tag{3}$$

where $H_t, \widehat{H}_t \in \mathbb{R}^{s\times d}$, and s, d indicates the length of the sequence and hidden dimension respectively.

Dialogue Modeling The representation $\widehat{H}_t$ obtained by the BART-Model is max-pooled to obtain the aggregated representation of the utterances as follows:

$$\check{h}_t = \text{max-pooling}(\widehat{H}_t). \tag{4}$$

To model the historical context information of the dialogue, we exploit a dialogue-level Transformer (Vaswani et al. 2017) layer as the context encoder. The multi-head attention mechanism can capture the interaction between different dialogues in multiple rounds of dialogue and aggregate different features to obtain the final implicit representation, thereby fully modeling the complex dependence between different utterances and context relations. For all utterances in a context, the multi-head attention score of the hidden state between two different utterances in a conversation $\check{h}_j$, $\check{h}_k$ can be calculated by the following formulas:

$$\text{Atten}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

$$head_i = \text{Atten}(\check{h}_j W_i^Q, \check{h}_k W_i^K, \check{h}_k W_i^V), \tag{6}$$

$$\text{MultiHead}(Q,K,V) = [head_1;\cdots;head_n]W^O, \tag{7}$$

where $W_i^Q \in \mathbb{R}^{d\times d_q}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W^O \in \mathbb{R}^{d\times d}$ are parameters that can be optimized, d_q, d_k and d_v are dimensions of query, key and value vectors, n indicates the number of heads.

Therefore, the utterance representation that models the context-dependence can be obtained through the above-mentioned dialogue-level Transformer:

$$H_{win} = [\check{h}_t, \check{h}_{t+1}, \cdots, \check{h}_{t+bs-1}], \tag{8}$$

$$H_{d\text{-}win} = \text{Dialogue-Transformer}(H_{win}), \tag{9}$$

where $H_{win} \in \mathbb{R}^{bs\times d}$ indicates utterances in a conversation within the window size bs and $H_{d\text{-}win} \in \mathbb{R}^{bs\times d}$ denotes the utterances after context modeling.

Supervised Contrastive Learning Supervised contrastive learning assumes that some crucial aspects get attention and allows few-shot learning to be more stable when fine-tuned on pre-trained models (Gunel et al. 2021). The typical contrastive learning uses only one pair of positive examples, while all other samples are treated as negative examples. Supervised contrastive learning treats all examples with the same label in the batch as positive examples by making full use of label information.

For ERC, the number of samples in each category in some datasets (Li et al. 2017) is highly unbalanced, while the supervised contrastive learning will mask itself when calculating the loss. If only one sample exists for a category in the batch, it cannot be directly applied to calculate the loss. Therefore, a copy of the hidden state of the utterance $H_{d\text{-}win}$ is made to obtain $\overline{H}_{d\text{-}win}$, and its gradient is detached. Hence the parameter optimization is maintained stable.

For a batch with N training samples, each sample is operated by the above mechanism to obtain multiview $2N$ samples, then the supervised contrastive loss of all samples in a batch can be expressed by the following equation:

$$X = [H_{d\text{-}win}, \overline{H}_{d\text{-}win}], \tag{10}$$

$$\mathcal{L}_{\text{SCL}} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \text{SIM}(p,i), \tag{11}$$

$$\text{SIM}(p,i) = \log \frac{\exp((X_i \cdot X_p)/\tau)}{\sum_{a\in A(i)} \exp(X_i \cdot X_a/\tau)}, \tag{12}$$

where $X \in \mathbb{R}^{2N\times d}$, $i \in I = \{1, 2, \cdots, 2N\}$ indicate the index of the samples in a multiview batch, $\tau \in R^+$ denotes the temperature coefficient used to control the distance between instances, $P(i) = I_{j=i} - \{i\}$ represents samples with the same category as i while excluding itself, $A(i) = I - \{i, N+i\}$ indicates samples in the multiview batch except itself.

Auxiliary Response Generation

To facilitate the model to consider richer contextual information when determining utterance sentiment, the model is required to generate its following utterance u_{t+1} given the current utterance u_t. The output hidden state of each token in u_{t+1} is generated by the BART decoder sequentially.

$$\acute{H}_t = \text{BART-Encoder}(H_t), \tag{13}$$

$$\grave{h}_j^d = \text{BART-Decoder}(\acute{H}_t; \grave{h}_{<j}^d), \tag{14}$$

$$u_{t+1,j} = \text{Softmax}(\grave{h}_j^d), \tag{15}$$

$$\mathcal{L}_{\text{Gen}} = -\sum_{i=1}^{N} \log p(u_{t+1}|u_t, \boldsymbol{\theta}), \tag{16}$$

where $\boldsymbol{\theta}$ is the parameters of BART need to be optimized.

Model Training

The loss of model training consists of three parts: the hidden state $H_{d\text{-}win}$ obtained after context modeling passes through a multilayer perceptron to obtain logits for calculating cross-entropy loss. The other part is the supervised contrastive loss and the loss of response generation. The loss is a weighted sum of the three components, and the sum of their weights equals one. The overall framework of CoG-BART is illustrated in Figure 2.

Dataset		DD	MELD	ENLP	IEMOCAP
#Dial	Train	11118	1038	713	120
	Dev	1000	114	99	120
	Test	1000	280	85	31
#Utter	Train	87170	9989	9934	5810
	Dev	8069	1109	1344	5810
	Test	7740	2610	1328	1623
#CLS		7	7	7	6

Table 1: Statistics of four benchmark datasets.

$$P_i = \mathrm{Softmax}(W_s H_{d\text{-}win,i} + b_s), \tag{17}$$

$$\hat{y}_i = \mathrm{argmax}(P_i), \tag{18}$$

$$L_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log \hat{y}_{i,c}, \tag{19}$$

$$\mathcal{L} = (1-\alpha-\beta)\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{\mathrm{SCL}} + \beta\mathcal{L}_{\mathrm{Gen}}, \tag{20}$$

where $y_{i,c}$ represents the label of a certain utterance, $\hat{y}_{i,c}$ indicates the probability distribution of category c output by the dense layer, α denotes the weight for supervised contrastive loss and β is the weight for loss of response generation.

Experimental Settings

This section will elaborate on the datasets, baseline models, experimental conditions, and parameter settings adopt in the experiment.

Experimental Setup

The code framework and initial weight of BART come from Huggingface's Transformers (Wolf et al. 2020). The optimizer applied for model training is AdamW with a linear-scheduled warm-up strategy. The parameters adjusted in this experiment include batch size, learning rate, warm-up ratio, α, and β. We conducted a hyperparameter search for model training through the reserved validation set. The results on the test set come from the best checkpoint in the validation set, and we average the scores from five different random seeds. All experiments are performed on GeForce RTX 3090 GPU.

Datasets

This section will introduce four benchmark datasets: MELD (Poria et al. 2019), EmoryNLP (Zahiri and Choi 2018), DailyDialog (Li et al. 2017), and IEMOCAP (Busso et al. 2008) for comparison with the baseline models.

MELD This dataset comes from the dialogue content of the characters in the American drama *Friends*. MELD originally contained multi-modal data, but we used only the text data for the experiments.

EmoryNLP (ENLP) This dataset also comes from *Friends*, and the difference from MELD is the annotation of utterance's emotional label category. The emotional tags contained in this dataset are: *joyful*, *neutral*, *powerful*, *mad*, *sad*, *scared*, and *peaceful*.

DailyDialog (DD) Manually compiled data sets about daily communication. The annotation method used in this data set is Ekman's emotion type (Ekman 1993), which includes six basic emotion tags, including *happiness*, *surprise*, *anger*, *disgust*, *fear*, and *sadness*.

IEMOCAP Like MELD, it is a multi-modal dataset. The content is derived from the lines in the scripts of the two actors, and the emotional tags included are *excited*, *neutral*, *frustrated*, *sad*, *happy*, and *angry*.

The detailed statistics of the four datasets are shown in Table 1, where "#Dial" indicates the number of dialogue in train/dev/test, "#Utter" represents the number of all utterances in dialogue, and "#CLS" denotes the number of categories of each dataset.

Metrics

For MELD, EmoryNLP and IEMOCAP, we adopt weighted average F1 as the evaluation metrics. In that "neutral" occupies the majority in DailyDialog, micro-F1 is employed as the evaluation metric for this data set, and we ignore the label "neutral" when calculating the results as in the previous works (Zhu et al. 2021; Shen et al. 2021b).

Results and Analysis

Main Results

Table 2 and 3 record the results of comparing CoG-BART with the baseline models on four datasets.

Among the pre-train-based models and their variants, the selected baseline models are BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), HiTrans (Li et al. 2020), DialogXL (Shen et al. 2021a) and XLNet (Yang et al. 2019). In MELD (Poria et al. 2019), CoG-BART has an approximate absolute 1.24% improvement over the previous state-of-the-art BART-large (Lewis et al. 2020).

For graph-based models, KET (Zhong, Wang, and Miao 2019), RGAT (Ishiwatari et al. 2020), DialogGCN (Ghosal et al. 2019), DialogCRN (Hu, Wei, and Huai 2021), COSMIC (Ghosal et al. 2020), and DAG-ERC (Shen et al. 2021b) are listed.

Compared to the graph-based model, CoG-BART improves 0.53 points over COSMIC (Ghosal et al. 2020). It is worth noting that RoBERTa-large is used as the feature extractor in COSMIC, while CoG-BART only adopts BART-large as the backbone structure to obtain competitive results, indicating that adequate knowledge transfer of pre-trained models which effectively model the dependencies between contexts can also yield promising results in MELD.

We can observe from the results in EmoryNLP (Zahiri and Choi 2018) that the graph-based model using the pre-trained model as the feature extractor works better overall than the model applying only the pre-trained model as the backbone network. Meanwhile, CoG-BART still achieves results with significant improvement. Also, the graph-based model can obtain higher F1 overall on IEMOCAP (Busso et al. 2008) compared to the pre-trained based models. The reason is that

Dataset	MELD		EmoryNLP		IEMOCAP		DailyDialog	
Model	Weighted -Avg-F1	Micro-F1	Weighted -Avg-F1	Micro-F1	Weighted -Avg-F1	Micro-F1	Weighted -F1-neural	Micro -F1-neutral
BERT	62.28	63.49	34.87	41.11	60.98	-	53.41	54.85
RoBERTa	62.51	63.75	35.90	40.81	63.38	-	52.84	54.33
HiTrans	61.94	-	36.75	-	64.50	-	-	-
DialogXL	62.41	-	34.73	-	65.94	-	-	54.93
XLNet	61.65	-	34.13	-	61.33	-	-	53.62
BART-large	63.57	64.41	35.98	38.93	56.14	56.67	54.83	55.34
CoG-BART	**64.81** (±0.19)	**65.95** (±0.44)	**39.04** (±0.10)	**42.58** (±0.94)	**66.18** (±0.45)	**66.71** (±0.49)	**56.09** (±0.01)	**56.29** (±0.17)

Table 2: The overall results of CoG-BART with pre-train-based baseline models on four datasets.

Dataset	MELD	EmoryNLP	IEMOCAP	DailyDialog
Model	Weighted -Avg-F1	Weighted -Avg-F1	Weighted -Avg-F1	Micro -F1-neutral
KET	58.18	34.39	59.56	53.37
RGAT	60.91	34.42	65.22	54.31
RGAT+RoBERTa	62.80	37.89	66.36	59.02
DialogGCN	58.10	-	64.18	-
DialogCRN	58.39	-	66.20	-
COSMIC	64.28	37.10	63.05	56.16
DAG-ERC	63.65	39.02	**68.03**	**59.33**
CoG-BART	**64.81** (±0.19)	**39.04** (±0.10)	66.18 (±0.45)	56.29 (±0.17)

Table 3: Comparison with graph-based models.

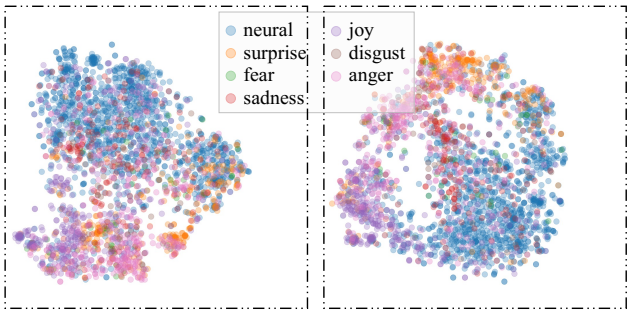

Figure 3: The t-SNE visualization results of the model output when α is 0 and 0.8, respectively.

the number of utterances contained in one context of IEMOCAP is much larger than the other three datasets, so pre-trained models are usually incapable of handling excessively long contexts. However, graph network models can better model context dependencies. In comparison, CoG-BART also achieves results similar to those of graph-based models, demonstrating the capability of CoG-BART to model the context-dependence.

The micro-F1 values of CoG-BART in DailyDialog are lower compared to the results of some graph neural network models. Still, it can achieve similar results to some pre-train-based models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019) and DialogXL (Shen et al. 2021a). Therefore, the graph-based model may have the advantage over pre-train-based models by more adequately modeling context dependencies on this dataset.

Metric	Weighted Average F1					
Datasets	α=0.2	α=0.4	α=0.6	α=0.8	β=0.1	β=0.2
MELD	**64.57**	63.99	64.42	61.84	64.83	63.70
IEMOCAP	64.38	**66.18**	65.12	63.38	66.18	63.54
EmoryNLP	**39.04**	36.68	36.90	35.24	37.45	37.57

Table 4: The F1 scores for different values of α and β

The Potency of Supervised Contrastive Learning

Qualitative Analysis of SCL To conduct a qualitative analysis of supervised contrastive learning, we utilize t-SNE (Hinton and Roweis 2002) to visualize the distribution of high-dimensional hidden states obtained by the model trained with supervised contrastive loss. By controlling different sizes of α, the ratio of supervised contrastive loss is controlled to 0% and 80%, respectively, to obtain the hidden state output by the model under different levels of supervised contrastive learning.

As illustrated in Figure 3, when the supervised contrastive loss is not exploited, that is, when the cross-entropy loss function is completely adopted, the overlap rate of samples between different labels is particularly high, especially for some samples with similar emotions, which increase the difficulty of learning the decision boundaries. As the proportion of supervised contrastive loss increases, it can be distinctly observed that the degree of coupling between different classes is gradually enlarged, and the same classes begin to cohesive. It is worth mentioning that although the distance within the class has been reduced, the uniformity (Wang and Isola 2020) between samples has been well maintained, indicating that the information has been well preserved and no representation collapse has occurred.

Quantitative Analysis of SCL The effects of different proportions of supervised contrastive learning on CoG-BART are illustrated in Table 4, where the weighted average F1 of CoG-BART with different proportions of SCL loss is recorded. Different α have a large impact on the outcomes, e.g., there exists a 2.8 points difference in F1 values between α equals 0.4 and 0.8 for IEMOCAP, reflecting the significant positive effect of supervised contrastive learning for this dataset. Meanwhile, different datasets have different values of α in obtaining the relatively best gain effect. For instance,

Utterance for Prediction	Generated Response	Predict w/o RG	Predict with RG	Golden label
Joey : Thursday's clearly not good for ya, pick a day!	Sarah : So that's two boxes of the Holiday Macaroons. On behalf of the Brown Birds of America, I salute you.	anger	joy	joy
Joey: Man, that was great! Huh? Can you believe how long we threw that ball around?	Rachel : Yeah, it is amazing it lasted that long.	surprise	joy	joy

Figure 4: Case studies show that response generation enables the model to correctly predict the emotion based on context.

Dataset	MELD	IEMOCAP
Methods	Weight-Avg-F1	
CoG-BART	64.81	66.18
-Gen	64.26 (↓0.55)	64.74 (↓1.44)
-SCL loss	64.28 (↓0.53)	64.23 (↓1.95)
-Speaker	64.14 (↓0.67)	55.41 (↓10.77)
-Gen, SCL loss	63.57 (↓**1.24**)	62.90 (↓3.28)
-SCL loss, Speaker	63.72 (↓1.09)	54.83 (↓**11.35**)
-Gen, Speaker	64.02 (↓0.79)	54.95 (↓11.23)
-Dialog-Trans	64.40 (↓0.41)	64.19 (↓1.99)

Table 5: Ablation study to evaluate the impact of different components on the overall performance of the model on MELD and EmoryNLP

CoG-BART performs best when $\alpha = 0.2$ in MELD, while achieving the best result when $\alpha = 0.4$ in IEMOCAP.

Effect of Response Generation

Response generation has a facilitating effect on modeling context dependence to some extent. As the two cases in Figure 4 illustrate, if only the current utterance itself is considered, the expression may cause the model to misjudge the sentiment of the current utterance, while generating responses leads the model to pay more attention to contextual information, thus making correct predictions which consistent with the scenario. As for the impact of different weights of response generation loss, Table 4 illustrates that when fixing α and adjusting β, there is also a slight impact on the model's overall performance.

Ablation Analysis

To investigate the impacts of individual modules and combinations of several components on the overall effect of the model, this section conducts an ablation study on three modules in CoG-BART. As illustrated in Table 5, the selected datasets are MELD and IEMOCAP, where "-" indicates the removal of the single method or several methods, "Gen" denotes the auxiliary task of response generation, "SCL loss" means supervised contrastive loss, and "Speaker" indicates the splicing of speaker label before utterance.

From the results of MELD, removing any of the three modules makes the overall performance worse, while discarding the supervised contrastive loss and response generation has the greatest impact on the performance of CoG-BART in MELD. These indicate that supervised contrastive loss leverage label information better compared to cross-entropy loss, thus effectively distinguishing different sentiments.

Consistent results are also obtained in IEMOCAP, indicating that the improvement in model performance from these three modules transfers well across these datasets. However, the more unexpected finding was that removing the speaker's information made CoG-BART most degraded in IEMOCAP. By analyzing this dataset, we found that it involved 302 speakers, so it may be crucial to fully model the speaker information for this dataset. It also proves the effectiveness of the simple method of splicing the speaker information directly in front of the utterance. Furthermore, removing the supervised contrastive loss alone degrades the performance by 1.95 points on IEMOCAP, indicating that supervised contrastive learning significantly impacts CoG-BART on this dataset. The results after removing Dialog-level Transformer suggest that this module improves overall performance by modelling longer contextual dependencies.

Conclusion

We propose supervised contrastive learning with response generation as an auxiliary task for BART, namely CoG-BART, for emotion recognition in conversation (ERC). First, supervised contrastive learning is introduced into the training process to distinguish similar emotions, reducing intra-class distance and increasing inter-class variance. Meanwhile, the response generation is adopted as an auxiliary task; hence, the model categorizes utterances with similar semantics but different emotions by considering the context. The results obtained on four datasets compared with the current state-of-the-art baseline methods demonstrate the proposed method's effectiveness. Furthermore, ablation studies demonstrate that supervised contrastive learning can effectively improve the model's efficacy in recognizing emotions, thus improving the overall performance. Also, response generation as an auxiliary task helps the model fully consider the context to discern the emotions of semantically similar utterances in varying contexts.

Acknowledgments

We are very grateful to the reviewers for their diligent and rigorous attitude towards our work and their valuable suggestions for improvement during the whole review process. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0108702) and the National Natural Science Foundation of China (NO. 62022027).

References

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4762–4779. Association for Computational Linguistics.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; and Sun, X. 2020a. Measuring and Relieving the Over-Smoothing Problem for Graph Neural Networks from the Topological View. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 3438–3445. AAAI Press.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ekman, P. 1993. Facial expression and emotion. *American psychologist*, 48(4): 384.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *CoRR*, abs/2104.08821.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481. Online: Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 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)*, 154–164. Hong Kong, China: Association for Computational Linguistics.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hinton, G. E.; and Roweis, S. T. 2002. Stochastic Neighbor Embedding. In Becker, S.; Thrun, S.; and Obermayer, K., eds., *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, 833–840. MIT Press.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 7042–7052. Association for Computational Linguistics.

Ide, T.; and Kawahara, D. 2021. Multi-Task Learning of Generation and Classification for Emotion-Aware Dialogue Response Generation. In Durmus, E.; Gupta, V.; Liu, N.; Peng, N.; and Su, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, NAACL-HLT 2021, Online, June 6-11, 2021*, 119–125. Association for Computational Linguistics.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 2020. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 7360–7370. Association for Computational Linguistics.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Kim, T.; Yoo, K. M.; and Lee, S. 2021. Self-Guided Contrastive Learning for BERT Sentence Representations. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 2528–2540. Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L.

2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Ji, D.; Li, F.; Zhang, M.; and Liu, Y. 2020. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4190–4200. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Lin, T.; Wang, Y.; Liu, X.; and Qiu, X. 2021. A Survey of Transformers. *arXiv preprint arXiv:2106.04554*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained Models for Natural Language Processing: A Survey. *SCIENCE CHINA Technological Sciences*, 63(10): 1872–1897.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 13789–13797. AAAI Press.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 1551–1560. Association for Computational Linguistics.

Sheng, D.; Wang, D.; Shen, Y.; Zheng, H.; and Liu, H. 2020. Summarize before Aggregate: A Global-to-local Heterogeneous Graph Inference Network for Conversational Emotion Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4153–4163. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 9929–9939. PMLR.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 5065–5075. Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 5754–5764.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Workshops*, 44–52. AAAI Press.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *CoRR*, abs/1909.10681.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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*, 1571–1582. Association for Computational Linguistics.