CEPT: a Contrast-Enhanced Prompt-Tuning Framework for Emotion Recognition in Conversation

Qingqing Gao, Jiuxin Cao*, Biwei Cao, Xin Guan, Bo Liu
School of Cyber Science and Engineering, Southeast University,
School of Computer Science and Engineering, Southeast University
No.2 Southeast University Road, Nanjing, Jiangsu, China
{qingqing_gao, jx.cao, caobiwei, xin_guan, bliu}@seu.edu.cn

Abstract

Emotion Recognition in Conversation (ERC) has attracted increasing attention due to its wide applications in public opinion analysis, empathetic conversation generation, and so on. However, ERC research suffers from the problems of data imbalance and the presence of similar linguistic expressions for different emotions. These issues can result in limited learning for minority emotions, biased predictions for common emotions, and the misclassification of different emotions with similar linguistic expressions. To alleviate these problems, we propose a Contrast-Enhanced Prompt-Tuning (CEPT) framework for ERC. We transform the ERC task into a Masked Language Modeling (MLM) generation task and generate the emotion for each utterance in the conversation based on the prompt-tuning of the Pre-trained Language Model (PLM), where a novel mixed prompt template and a label mapping strategy are introduced for better context and emotion feature modeling. Moreover, Supervised Contrastive Learning (SCL) is employed to help the PLM mine more information from the labels and learn a more discriminative representation space for utterances with different emotions. We conduct extensive experiments and the results demonstrate that CEPT outperforms the state-of-the-art methods on all three benchmark datasets and excels in recognizing minority emotions.

Keywords: Emotion recognition in conversation, Prompt-tuning, Contrastive learning

1. Introduction

Emotion recognition in conversation (ERC) is an emerging research area in Natural Language Processing (NLP) that aims to recognize the emotion of each utterance in a conversation.

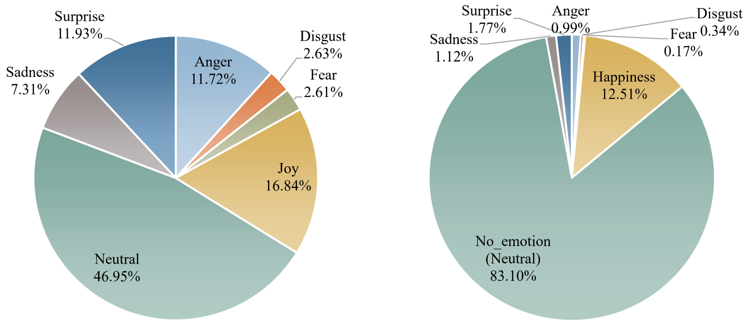
Different from the traditional sentence-level emotion recognition, ERC faces the challenge that the emotion expressed by each utterance is influenced by both its own semantics and contextual factors, including adjacent utterances and speakers (Ghosal et al., 2019). Moreover, people may use similar linguistic expressions to convey different emotions, such as anger and surprise, which makes it challenging to detect the subtle differences (Li et al., 2022). Furthermore, people tend to remain calm during most conversations and only express strong emotions, like disgust or fear, in some particular situations (Jiao et al., 2019). Thus, imbalanced emotion distributions are prevalent in ERC datasets, as shown in Figure 1 illustrating the emotion distributions in three benchmark datasets: MELD (Poria et al., 2019), DailyDialog (Li et al., 2017) and IEMOCAP (Busso et al., 2008). This can easily lead to limited learning for minority emotions and biased predictions for common emotions.

Recent research on ERC has focused on leveraging the advancements of Pre-trained Language Models (PLMs) (Shen et al., 2021a; Gao et al., 2022; Shen et al., 2021b; Li et al., 2022). Scholars commonly fine-tune PLMs with ERC datasets to adapt them for ERC, but a gap exists between pre-training objectives (e.g., Masked Language Modeling, MLM) and ERC. Fine-tuning PLMs with high-quality labeled data and sufficient training epochs is crucial, but limited samples for minority emotions present challenges. Recently, prompt-tuning has gained attention for its ability to improve PLM performance on downstream tasks (**?**Wu et al., 2022; Yang et al., 2023) and has become a new paradigm in modern NLP (Ding et al., 2022). The idea behind prompt-tuning is to transform the downstream task into a form resembling a pre-training objective and provide task descriptions, also known as prompts, to help the PLM to comprehend the task more effectively. Unlike fine-tuning, prompt-tuning enables efficient adaptation of PLMs to specific downstream tasks with fewer training samples and less training time, benefiting from the direct leverage of PLMs' extensive knowledge of linguistic patterns and structures from large corpora. This can help alleviate the issue of limited labeled data for minority emotions. Nevertheless, implementing a prompt-tuning framework for ERC requires careful consideration of how to effectively model contextual information and accurately distinguish different emotions when their linguistic expressions are similar.

To alleviate the challenges discussed, we convert the ERC problem into an MLM problem, applying the PLM's MLM capability through prompt-tuning to generate emotions for utterances in conversations to mitigate the data imbalance issue. Moreover,

*Corresponding author.

LREC-COLING 2024, pages 2947–2957



(a) Emotion distribution in MELD. (b) Emotion distribution in DD. (c) Emotion distribution in IEMOCAP.

Figure 1: Emotion distributions in three benchmark datasets. The abbreviation "DD" refers to DailyDialog.

we design a context-aware mixed prompt template to effectively model the conversation context and a label mapping strategy for enhanced emotion modeling. Furthermore, Supervised Contrastive Learning (SCL) is employed to improve the PLM's ability to better distinguish utterances with different emotions.

The main contributions of this work are as follows:

- We propose a novel Contrast-Enhanced Prompt-Tuning (CEPT) framework for ERC. CEPT transforms ERC into an MLM problem to generate emotions for utterances in conversations. By bridging the gap between PLM's MLM and ERC, CEPT can utilize PLM's inherent linguistic knowledge from large corpora more effectively, mitigating the challenges posed by imbalanced data. Furthermore, CEPT incorporates SCL to mine more information from emotion labels and learn more discriminative representations for utterances with different emotions.

- A context-aware mixed prompt template is designed that integrates both hard words and soft words. Hard words indicate context and target utterance boundaries, while adjustable soft words assist the PLM in learning a customized prompt that effectively guides the generation of appropriate emotion words. Moreover, we introduce a label mapping strategy for comprehensive emotion modeling.

- CEPT is evaluated on three benchmark datasets and it consistently outperforms state-of-the-art methods. Furthermore, CEPT demonstrates excellent performance in recognizing minority emotions. Additionally, extensive experiments are conducted to thoroughly evaluate the effectiveness of CEPT.

2. Related work

2.1. Emotion recognition in conversation

In the early stages, recurrence-based models are extensively used for ERC (Hazarika et al., 2018b,a; Majumder et al., 2019; Jiao et al., 2019; Hu et al., 2021) due to their ability to capture both temporal and semantic information. However, recurrence-based models can struggle with capturing long-range contextual information. In recent years, significant advancements in PLMs have sparked their widespread adoption for ERC (Shen et al., 2021b,a; Gao et al., 2022). Although models based on PLMs have shown promising results in ERC, the gap between ERC and the pre-training objectives means it needs enough high-quality labeled data and training epochs to fine-tune the PLMs. Limited samples for minority emotions in ERC pose challenges for PLM-based methods to achieve a good performance.

2.2. Prompt-tuning

Prompt-tuning aligns the downstream task with PLM pre-training objectives by modifying the input with a prompt template. Early research focuses on fixed hard templates (Petroni et al., 2019; Gao et al., 2021a; Schick and Schütze, 2021), which lack flexibility. Therefore, some works explore prompts with learnable vectors such as prefix-tuning (Li and Liang, 2021; Liu et al., 2022), soft prompt (Lester et al., 2021; Wu and Shi, 2022), P-tuning (Liu et al., 2021b) and P-tuning V2 (Liu et al., 2021a). For ERC, CISPER (Yi et al., 2022) generates continuous prompt embeddings based on semantic features and commonsense knowledge, and simply concatenates the embeddings with a "mask", which lacks prompt interpretability and relies on external knowledge. PLMs, trained on large corpora, inherently possess commonsense knowledge. Through effective utilization of PLMs, the need for external knowledge sources can be eliminated.

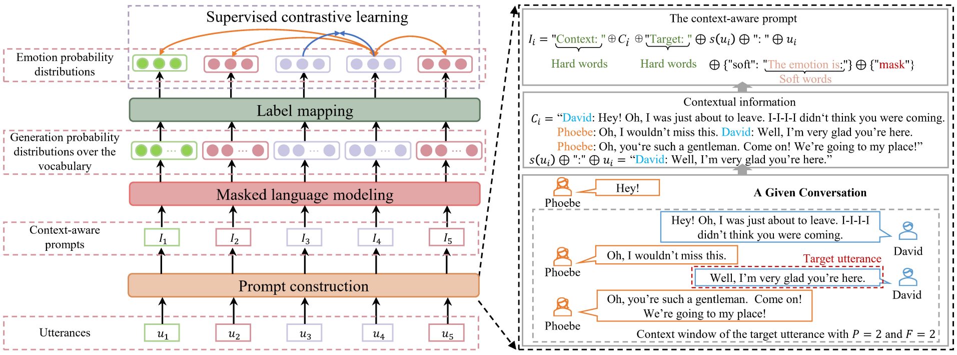

Supervised contrastive learning
Emotion probability distributions
Label mapping
Generation probability distributions over the vocabulary
Masked language modeling
Context-aware prompts
I_1
I_2
I_3
I_4
I_5
Prompt construction
Utterances
u_1
u_2
u_3
u_4
u_5
The context-aware prompt
I_i = "Context: " ⊕ C_i ⊕ "Target: " ⊕ $s(u_i)$ ⊕ ": " ⊕ u_i
Hard words
Hard words
⊕ {"soft": "The emotion is:"} ⊕ {"mask"}
Soft words
Contextual information
C_i = "David: Hey! Oh, I was just about to leave. I-I-I-I didn't think you were coming.
Phoebe: Oh, I wouldn't miss this. David: Well, I'm very glad you're here.
Phoebe: Oh, you're such a gentleman. Come on! We're going to my place!"
$s(u_i)$ ⊕ ":" ⊕ u_i = "David: Well, I'm very glad you're here."
Phoebe
Hey!
A Given Conversation
Hey! Oh, I was just about to leave. I-I-I-I didn't think you were coming.
David
Phoebe
Oh, I wouldn't miss this.
Target utterance
Well, I'm very glad you're here.
David
Phoebe
Oh, you're such a gentleman. Come on! We're going to my place!
Context window of the target utterance with $P = 2$ and $F = 2$

Figure 2: The architecture of CEPT. It comprises four key components: prompt construction, MLM, label mapping, and SCL. The example in a black dashed box shows the prompt construction process. For the SCL part, the yellow arrows pointing opposite and the blue arrows pointing toward represent repulsion and aggregation, respectively.

2.3. Contrastive learning

Contrastive learning maximizes similarity between similar instances and minimizes similarity between dissimilar instances, making it effective for learning representations from unlabeled data (He et al., 2020; Chen et al., 2020). Khosla et al. (2020) extend this approach to a supervised version called Supervised Contrastive Learning (SCL) by grouping samples with the same label together and pushing samples with different labels apart in the embedding space. Gunel et al. (2021) propose leveraging SCL to guide the fine-tuning process of PLMs. In the field of ERC, CoG-BART (Li et al., 2022) incorporates SCL to assist the fine-tuning of BART (Lewis et al., 2020). Moreover, SPCL-CL-ERC (Song et al., 2022) propose a supervised prototypical contrastive learning loss to guide the SimCSE (Gao et al., 2021b). Motivated by these studies, we propose leveraging SCL to enhance the prompt-tuning of PLM.

3. Methodology

3.1. Problem definition

Given a conversation with N utterances $U = [u_1, u_2, ..., u_N]$ uttered by a sequence of speakers $S = [s(u_1), s(u_2), ..., s(u_N)]$, where s maps the utterance into the corresponding speaker, the ERC task is to recognize the emotion label y_i of the utterance u_i. The emotion label y_i belongs to the predefined emotion category set $E = [e_1, e_2, ..., e_M]$ and M is the number of the pre-defined emotion categories.

3.2. Architecture

The architecture of CEPT is shown in Figure 2. Specifically, we first design a mixed prompt template for prompt construction. Each utterance, along with its contextual information, is modified using the prompt template to obtain the prompt. The prompts are then input to the PLM, which performs MLM to predict generation probability distributions over the vocabulary. Label words associated with predefined emotion categories are selected from the vocabulary, and the PLM's generation probabilities for each label word are integrated to obtain probability distributions over emotion categories. Finally, SCL helps the PLM to learn a representation space where the utterances with the same emotion are clustered together and those with different emotions are separated apart.

3.3. Prompt construction

We design a context-aware mixed prompt template, combining both hard words and soft words, for prompt construction. Specifically, We set two hard words, which are "Context" and "Target", to indicate the boundaries of context and target utterance. Meanwhile, we introduce soft words, initialized as "The emotion is: " to guide the PLM to generate the emotion of the target utterance. These soft words are adjusted during prompt-tuning to allow the model to learn a prompt template that optimally supports the ERC task. Additionally, the "mask" keyword is indispensable as it directs the PLM to predict the missing emotion word. The prompt template is formalized as follows to combine both the contextual and semantic information of the target utterance u_i:

$$\begin{aligned} I_i =& ”Context: ” \oplus C_i \\ & \oplus ”Target: ” \oplus s(u_i) \oplus ” : ” \oplus u_i \\ & \oplus \{”soft” : ”The\, emotion\, is : ”\} \\ & \oplus \{”mask”\}, \end{aligned} \tag{1}$$

$$\begin{aligned} C_i =& [s(u_{i-P}) : u_{i-P},\ s(u_{i-P+1}) : u_{i-P+1}, \\ & \ldots, s(u_i) : u_i, \ldots, s(u_{i+S-1}) : u_{i+S-1}], \end{aligned} \tag{2}$$

Category	Label words
Anger / Angry	anger, frustration, irritation, hostility
Disgust	disgust, revulsion, nausea, aversion
Fear	fear, anxiety, apprehension, nervousness
Happiness / Joy / Happy	joy, happiness, contentment, satisfaction
Sadness / Sad	sadness, loss, disappointment, grief
Surprise	surprise, astonishment, amazement
Excited	excitement, thrill, exhilaration
Frustrated	frustration, disappointment, dissatisfaction, annoyance
Neutral / no_ emotion	neutral / no_ emotion

Table 1: Mapping between original emotion category and label words.

where $\oplus$ represents concatenation, I_i is the prompt for u_i and "mask" donates the masked position. C_i is the contextual information of u_i, which comes from both the nearby utterances and speakers. To avoid information overload and excessive computational requirements, we set a context window to limit the context range. The size of the window is set as $P + S$, where P and S represent the number of past and succeeding utterances from u_i respectively. Each utterance is modified using this template to construct the corresponding prompt and the prompts are fed into the PLM for MLM to predict the generation probability distributions over the vocabulary for the masked positions.

3.4. Label mapping

To enhance PLMs' understanding of the emotions, we map the original emotion category e_j to its label words set $EW_j = \{ew_1, ew_2, ..., ew_{k_j}\}$ as shown in Table 1, where k_j is the number of the label words corresponding to the emotion category e_j. The emotion categories in Table 1 are sourced from the MELD (Poria et al., 2019), DailyDialog (Li et al., 2017), and IEMOCAP (Busso et al., 2008) datasets. The label words selected for each emotion category include the original emotion category noun, synonyms of the original emotion category, and other words that convey related feelings. The generation probabilities of the label words predicted by the PLM are extracted, and the probability that the emotion category of utterance u_i is e_j, denoted as $p(e_j|u_i)$, is calculated as follows:

$$p(e_j|u_i) = \sum_{ew_j \in EW_j} p([MASK] = ew_j|I_i), \tag{3}$$

where $j \in \{1, 2, ..., M\}$.

The predicted emotion probability distribution over all the pre-defined emotion categories of utterance u_i is denoted as $\mathbf{P_i} \in R^M$. For a batch with D utterances, the loss of MLM generation is calculated using cross-entropy loss as follows:

$$L_{Gen} = -\frac{1}{D}\sum_{i=1}^{D}[\mathbf{y_i}]log(\mathbf{P_i}), \tag{4}$$

where $[\mathbf{y_i}]$ represents the one-hot vector of the ground truth emotion label of the utterance u_i.

3.5. Supervised contrastive learning

We employ SCL to enhance the prompt-tuning process, where examples with the same label within a batch are considered positive examples while examples with different labels serve as negative examples. Due to the common existence of data imbalance in ERC, a certain emotion label may only appear once in a batch, making it impossible to compute similarity directly. Motivated by CoG-BART (Li et al., 2022), we stack the predicted emotion probability distribution vectors of all utterances within a batch as a matrix H_b and make a copy of H_b, denoted as H_b', whose gradient is detached to ensure the parameter optimization is stable. The vectors used for computing the SCL loss are denoted as $H^m = [H_b, H_b'] = \{h_1^m, h_2^m, ..., h_D^m, h_{D+1}^m, ..., h_{2D}^m\}$ and the calculation formula for SCL loss is given as follows:

$$SIM(h_i^m, h_p^m) = \log \frac{exp(h_i^m \cdot h_p^m)/\tau)}{\sum_{a \in A(i)} exp(h_i^m \cdot h_a^m)/\tau)}, \tag{5}$$

$$L_{SCL} = \sum_{i=1}^{2D} \frac{-1}{|P(i)|} \sum_{p \in P(i)} SIM(h_i^m, h_p^m), \tag{6}$$

where $SIM(h_i^m, h_p^m)$ represents the similarity between h_i^m and h_p^m, τ is a scalar temperature parameter to control the sensitivity of the similarity calculation, $A(i)$ denotes the indices of the vectors in the H^m except i and $D + i$, and $P(i)$ represents the indices of the vectors corresponding to the utterances that share the same ground truth emotion label as the utterance u_i^m while excluding i.

3.6. Training

The loss that guides the training process of CEPT is a weighted combination of the MLM generation loss and the SCL loss:

$$L = (1 - \alpha)L_{Gen} + \alpha L_{SCL}, \quad (7)$$

where α is a hyperparameter that denotes the weight for SCL loss while $(1 - \alpha)$ is the weight for MLM generation loss.

4. Experimental settings

4.1. Dataset

We employ three benchmark datasets, offering diverse conversational contexts, to evaluate our framework CEPT.

MELD (Poria et al., 2019). It contains multi-modal multi-party conversations from the TV series *Friends*. The utterances are annotated with one of the seven emotion labels, which are anger, disgust, fear, joy, sadness, surprise, and neutral.

DailyDialog (Li et al., 2017). It consists of human-written dyadic conversations about daily life. The dataset also provides annotations for seven emotion labels, which are anger, disgust, fear, happiness, sadness, surprise, and no_emotion.

IEMOCAP (Busso et al., 2008). It is a multi-modal dyadic conversation dataset recorded from ten actors. The utterances are annotated with one of the six emotion labels, which are angry, happy, sad, excited, frustrated, and neutral.

We only use the textual data from the aforementioned datasets for our experiments, following Shen et al. (2021a,b); Gao et al. (2022); Li et al. (2022).

4.2. Evaluation metrics

We evaluate the performance on MELD and IEMOCAP using weighted average F1. For DailyDialog, we use micro average F1 and exclude the "Neutral" labels due to their overabundance, following the previous studies(Shen et al., 2021a,b; Gao et al., 2022; Li et al., 2022). In addition, we use the F1 score to assess the performance of each method under each emotion category.

4.3. Compared methods

DialogueRNN (Majumder et al., 2019) uses a Convolutional Neural Network (CNN) to extract utterance features and three Gated Recurrent Units (GRUs) to model the context and the speakers, and decoder the emotions.

DialogXL (Shen et al., 2021a) improves the XLNet (Yang et al., 2019) with enhanced memory and dialog-aware self-attention for ERC.

DAG-ERC (Shen et al., 2021b) is a graph-based method that uses the RoBERTa (Liu et al., 2019) to encode utterances and models the context using a directed acyclic graph.

ESD-ERC (Gao et al., 2022) uses BERT (Devlin et al., 2019) as the utterance encoder and use a Transformer encoder (Vaswani et al., 2017) to model the context with an emotion shift detection auxiliary task.

CoG-BART (Li et al., 2022) combines the BART (Lewis et al., 2020) with SCL and uses response generation as the auxiliary task.

CISPER (Yi et al., 2022) uses continuous prompts based on semantic features and external commonsense knowledge to prompt-tune the RoBERTa (Liu et al., 2019) for ERC.

SPCL-CL-ERC (Song et al., 2022) uses SimCSE (Gao et al., 2021b) for context encoding with a supervised prototypical contrastive learning loss and use a curriculum learning (Bengio et al., 2009) strategy to reorganize the datasets.

RoBERTa-ERC is the RoBERTa (Liu et al., 2019) with a linear layer and a softmax layer on the top for ERC.

CEPT is our proposed framework and use RoBERT (Liu et al., 2019) as the backbone.

4.4. Other experimental settings

For CEPT, we utilize the default hyper-parameters of roberta-large, with a seed value of 777, a batch size of 4, and a learning rate of 10^{-6}. Training epochs are 3 for MELD and 4 for DailyDialog and IEMOCAP. Other hyper-parameters are optimized using validation data. For other methods, we present either the reported results or the outcomes obtained using the provided code. In all experiments, the model with the best performance on the validation set is used for test evaluation.

5. Result analysis

5.1. Overall performance

Dataset	MELD	DD	IEMOCAP
DialogueRNN	57.10	50.27	62.75
ESD-ERC	62.15	57.44	-
DialogXL	62.67	54.93	65.94
DAG-ERC	63.42	59.33*	68.03
CoG-BART	64.90	56.29	66.18
CISPER	66.08	-	-
SPCL-CL-ERC	66.96*	-	69.74*
RoBERTa-ERC	64.61	52.87	51.65
CEPT	**67.51**	**61.52**	**70.53**

Table 2: Performance Comparison with the baseline and state-of-the-art methods. The best performances are in bold font and the second-best performances are marked with asterisks (*).

Emotion (Number)	Anger (345)	Disgust (68)	Fear (50)	Joy (402)	Neutral (1256)	Sadness (208)	Surprise (281)
DialogueRNN	42.26	00.00	00.00	52.79	76.11	21.59	46.78
ESD-ERC	48.40	00.00	00.00	59.49	79.21	27.33	58.41
DialogXL	49.93	00.00	00.00	61.25	78.55	33.16	57.56
DAG-ERC	49.17	30.09*	26.98	60.25	77.22	36.57	58.22
CoG-BART	47.34	19.35	30.00*	62.15	79.47	**43.40**	58.41
CISPER	56.80	23.53	28.89	61.37	80.53	38.83	56.69
SPCL-CL-ERC	56.91*	27.66	25.88	63.34*	80.57*	42.01	58.98*
RoBERTa-ERC	50.74	24.00	9.84	61.89	79.53	39.13	57.28
CEPT	**57.06**	**32.32**	**31.58**	**64.53**	**80.73**	42.39*	**59.02**

Table 3: F1 scores under each category of each model on MELD. The best performances are in bold font and the second best performances are marked with asterisks (*).

Prompt	Label mapping	SCL	MELD (Weighted-F1)	DailyDialog (micro-F1)	IEMOCAP (Weighted-F1)
$\surd$	$\surd$	$\surd$	67.51	61.52	70.53
$\times$	$\surd$	$\surd$	65.75 (↓1.76)	55.18 (↓6.34)	51.85 (↓18.68)
$\surd$	$\times$	$\surd$	66.29 (↓1.22)	57.66 (↓3.86)	68.31 (↓2.22)
$\surd$	$\surd$	$\times$	66.02 (↓1.49)	60.96 (↓0.56)	67.88 (↓2.65)

Table 4: The ablation results of CEPT on three datasets.

	Template
(1)	$''Context:'' \oplus C_i \oplus ''Target:'' \oplus s(u_i) \oplus '':'' \oplus u_i \oplus ''The\,emotion\,is:'' \oplus \{''mask''\}$
(2)	$\{''soft''\} \oplus C_i \oplus \{''soft''\} \oplus s(u_i) \oplus '':'' \oplus u_i \oplus \{''soft''\}\{''soft''\}\{''soft''\} \oplus \{''mask''\}$
(3)	$\{''soft'':''Context:''\} \oplus C_i \oplus \{''soft'':''Target:''\} \oplus s(u_i) \oplus '':'' \oplus u_i$ $\oplus\{''soft'':''The\,emotion\,is:''\} \oplus \{''mask''\}$
(4)	$''Context:'' \oplus C_i \oplus ''Target:'' \oplus s(u_i) \oplus '':'' \oplus u_i \oplus \{''soft''\}\{''soft''\}\{''soft''\} \oplus \{''mask''\}$
(5)	$''Context:'' \oplus C_i \oplus ''Target:'' \oplus s(u_i) \oplus '':'' \oplus u_i \oplus \{''soft'':''The\,emotion\,is:''\} \oplus \{''mask''\}$

Table 5: The templates with different strategies.

The performance comparison of the compared methods is presented in Table 2. The abbreviation "DD" refers to the DailyDialog dataset.

For MELD, CEPT surpasses the second-best by 0.55%. CEPT also outperforms CISPER, which incorporates external knowledge, by a margin of 1.43%. This highlights the effectiveness of CEPT in leveraging the inherent knowledge of the PLM. For DailyDialog, CEPT outperforms the second-best by 2.19% in terms of micro-F1 score. For IEMOCAP, CEPT surpasses the second-best by 0.79%. Additionally, we train Reborta for twice the number of epochs compared to CEPT but Reborta still significantly underperforms, further showing CEPT's superior utilization of the PLM.

5.2. Performance on each emotion category

Table 3 presents the performance comparisons in each emotion category of different methods on MELD. CEPT demonstrates superior performance compared to other methods in most emotion categories and achieves the second-best performance in the remaining category Sadness. CoG-BART surpasses CEPT in Sadness but obviously underperforms in other emotion categories.

Notably, CEPT shows particularly outstanding results in Disgust and Fear, surpassing the second-best by 2.23% and 1.58% respectively. Disgust and Fear are challenging to recognize due to the extremely limited number of available samples, resulting in zero weighted-F1 scores for the first three methods. However, our model demonstrates excellent performance in recognizing these two emotions, showing its outstanding ability to capture more information for minority emotions.

5.3. Ablation study

We conduct ablation experiments to further study the effectiveness of each component in CEPT. Table 4 shows the results, where the "$\times$" denotes the removal of a component, "$\surd$" denotes the retention of a component.

The results demonstrate that removing any of the three components leads to a decline in CEPT's performance. Specifically, the prompt component has the most significant impact, as its removal leads to the largest decrease in the weighted-F1 score across all three datasets compared to the removal of label mapping and the removal of SCL. This highlights the crucial role of the prompt template we design. Additionally, both the label mapping component and the SCL component can also significantly impact the model's performance.

5.4. Prompt analysis

We explore various prompt construction strategies and evaluate the impact of different prompt templates.

	MELD	DailyDialog	IEMOCAP
(1)	64.97	60.60	65.34
(2)	65.54	60.06	66.91
(3)	66.08	59.48	64.73
(4)	64.27	60.72	65.32
(5)	**67.51**	**61.52**	**70.53**

Table 6: The performance of CEPT with different prompt templates. The best performances are highlighted in bold font.

Specifically, we try five strategies, including: (1) hard template, (2) soft template without initialization, (3) soft template with initialization, (4) mixed template with hard words and uninitialized soft words, (5) mixed template with hard words and initialized soft words. The five prompt templates using five different strategies are shown in Table 5. Table 6 illustrates the performance of CEPT with different prompt templates. CEPT achieves the best performance when using the mixed template with hard words and initialized soft words. The inclusion of hard words provides clear indications of the context and target utterance boundaries, which have straightforward meanings, eliminating any potential noise or confusion. Additionally, the introduction of soft words, initialized as "The emotion is," serves as an initial guide for the PLM to generate the emotion of the target utterance. These soft words are adjustable, enabling the PLM to learn a customized ERC prompt template.

5.5. Parameters analysis

Context window size. We analyze the impact of the context window size on the performance of our CEPT framework across three datasets. We manipulate two parameters: P, representing the number of past utterances (ranging from 5 to 10), and S, representing the number of succeeding utterances (ranging from 3 to 8). The experimental results are presented in Tables 7-9. For MELD, the highest performance is achieved with $P = 8$ and $S = 4$. For DailyDialog, the optimal performance is obtained with $P = 8$ and $S = 7$. For IEMOCAP, the best performance is observed with $P = 9$ and $S = 5$. An overly small window size limits the model's ability to capture sufficient contextual information, while an excessively large window size introduces unnecessary noise. The most effective range of contextual information mainly depends on both the conversation length and the utterance length. As a result, the optimal context window size varies for each dataset due to differences in average conversation length and average utterance length.

S P	3	4	5	6	7	8
10	67.04	66.64	67.59	66.81	66.71	66.87
9	66.93	66.89	66.81	66.71	67.21	66.52
8	66.74	**67.81**	67.32	67.05	66.68	67.14
7	66.67	67.53	67.51	67.11	66.65	66.14
6	66.37	67.33	66.96	66.81	66.71	66.81
5	66.18	67.03	67.48	67.16	66.26	65.98

Table 7: The weighted-F1 results of CEPT on MELD with different context window sizes. Bold font denotes the best performance.

S P	3	4	5	6	7	8
10	58.61	58.64	61.00	58.91	60.64	59.99
9	59.16	60.44	58.76	59.14	60.28	59.54
8	60.13	60.36	59.63	59.80	**61.52**	60.38
7	60.87	60.66	59.30	61.10	59.80	59.45
6	59.59	60.89	59.06	60.00	59.68	60.20
5	60.44	59.30	59.85	58.47	57.86	59.14

Table 8: The weighted-F1 results of CEPT on DailyDialog with different context window sizes. Bold font denotes the best performance.

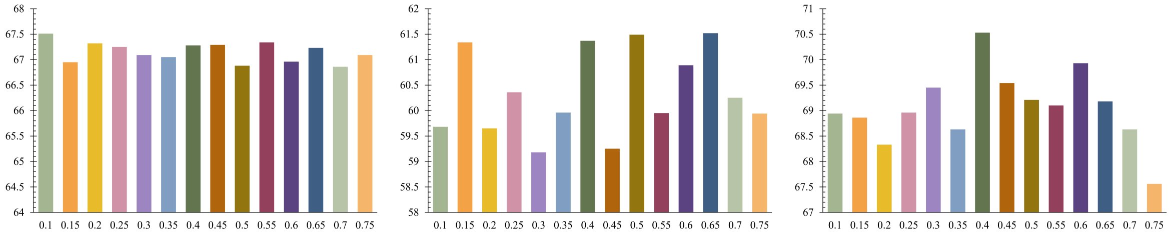
S P	3	4	5	6	7	8
10	68.50	69.31	68.30	67.59	68.19	67.97
9	68.21	67.70	**70.53**	67.82	68.27	68.09
8	69.03	69.04	69.96	69.21	69.24	69.69
7	67.50	66.68	68.74	70.35	67.81	67.32
6	66.74	68.78	67.77	67.03	67.88	67.50
5	66.51	67.83	66.88	67.42	67.22	67.79

Table 9: The weighted-F1 results of CEPT on IEMOCAP with different context window sizes. Bold font denotes the best performance.

SCL loss weight. We conduct experiments with different SCL loss weights, denoted as α, and present the results in Table 10 and figure 3. For MELD, the

α	0.1	0.15	0.2	0.25	0.3	0.35	0.4
MELD	**67.51**	66.95	67.32	67.25	67.09	67.05	67.28
DailyDialog	59.68	61.34	59.65	60.36	59.18	59.96	61.37
IEMOCAP	68.94	68.86	68.33	68.96	69.45	68.63	**70.53**
α	0.45	0.5	0.55	0.6	0.65	0.7	0.75
MELD	67.29	66.88	67.34	66.96	67.23	66.86	67.09
DailyDialog	59.25	61.49	59.95	60.89	**61.52**	60.25	59.94
IEMOCAP	69.54	69.21	69.10	69.93	69.18	68.63	67.56

Table 10: The results of CEPT with different weights of the supervised contrastive loss. Bold font denotes the best performance.



(a) The results of CEPT on MELD with different weights of the supervised contrastive loss.

(b) The results of CEPT on DailyDialog with different weights of the supervised contrastive loss.

(c) The results of CEPT on IEMOCAP with different weights of the supervised contrastive loss.

Figure 3: Visual comparison of CEPT performance with different supervised contrastive loss weights.

best performance is achieved with $\alpha = 0.1$, indicating the dominance of the cross-entropy loss. Similarly, for IEMOCAP, $\alpha = 0.4$ yields optimal performance, cross-entropy loss remains the dominant factor. Conversely, CEPT achieves optimal performance on DailyDialog with an SCL weight of $\alpha = 0.65$, indicating the dominant influence of the SCL loss compared to the cross-entropy loss. The varying impact of SCL loss weight on CEPT can be attributed to the dataset characteristics. MELD and IEMOCAP, sourced from actors, often exhibit magnified and exaggerated linguistic expressions, making the reliance on SCL less crucial. In contrast, the DailyDialog dataset consists of conversational data from everyday life, requiring a stronger emphasis on SCL to capture subtle emotions. Figure 3 visually demonstrates the varying influence of SCL weights on different datasets. MELD demonstrates less sensitivity to SCL weights, while DailyDialog and IEMOCAP show higher sensitivity.

5.6. Case Study

We analyze two cases from MELD depicted in Tables 11 and 12, where CEPT outperforms CISPER and SPCL-CL-ERC models.

For the case from Table 11, CISPER's first error might be attributed to the linguistic similarities between the third utterance and expressions commonly associated with Anger. In contrast, CEPT, with its prompt-tuning enhanced by SCL, demonstrates an exceptional ability to differentiate between different emotions that have similar linguistic expressions. CISPER's second error may be linked to the high prevalence of Neutral emotions in MELD, as evident in Figure 1. The abundance of Neutral samples can cause models to have a tendency to predict Neutral emotions, while CEPT avoids it. This shows CEPT's ability to prevent biased predictions for common emotions, alleviating imbalanced data issues.

For the case from Table 12, the misclassification by SPCL-CL-ERC may be attributed to insufficient learning of the Disgust emotion with very few samples. This emphasizes the superior performance of CEPT in mining more information for the minority emotions, mitigating the challenges posed by the imbalanced emotion distributions.

Overall, the presented two cases provide evidence that CEPT can effectively tackle two key challenges for ERC: the presence of similar linguistic expressions conveying different emotions and imbalanced emotion distributions, empowering it to achieve state-of-the-art results in ERC.

6. Conclusion

Our proposed CEPT transforms the Emotion Recognition in Conversation (ERC) task into a Masked Language Modeling (MLM) generation problem. This approach bridges the gap between

Speaker	Utterance	Ground truth	CEPT	CISPER
Ross	Hi.	Neutral	Neutral	Neutral
Rachel	Rachel.	Neutral	Neutral	Neutral
Ross	Rachel! Well, you-you're not at home, you're-you're-you're right here.	Surprise	Surprise	**Anger**
Rachel	Yeah I know, and I bet you thought it would be weird. But it's not!	Joy	Joy	**Neutral**
Ross	Okay. So well I'll umm, I'll have her home by midnight.	Neutral	Neutral	Neutral

Table 11: A case from MELD with the ground-truth emotion labels and the predicted labels from CEPT and CISPER. CISPER incorrectly classifies the emotion of the third utterance as Anger instead of Surprise and the emotion of the fourth utterance as Neutral instead of Joy, while CEPT accurately recognizes the emotions. The misjudgments are highlighted in bold.

Speaker	Utterance	Ground truth	CEPT	SPCL-CL-ERC
Joey	Here. I need to borrow some moisturizer.	Neutral	Neutral	Neutral
Monica	For what?	Neutral	Neutral	Neutral
Joey	Whaddya think? Today's the big day!	Joy	Joy	Joy
Monica	Oh my God. Okay, go into the bathroom, use whatever you want, just don't ever tell me what you did in there.	Disgust	Disgust	**Surprise**
Joey	Thank you!	Joy	Joy	Joy

Table 12: A case from MELD with the ground-truth emotion labels and the predicted labels from CEPT and SPCL-CL-ERC. SPCL-CL-ERC incorrectly classifies the emotion of the fourth utterance as Surprise instead of Disgust, while CEPT accurately recognizes it. The misjudgment is highlighted in bold.

PLM's MLM and ERC, reducing the need for extensive labeled data to tune the PLM. Moreover, CEPT introduces a context-aware mixed prompt template and a label mapping strategy, enhancing the PLM's ability to capture contextual information and various emotional features. Furthermore, Supervised Contrastive Learning (SCL) is employed to enable the PLM to extract more information from the labels and learn a more discriminative representation space for utterances with different emotions. Experiment results demonstrate that CEPT outperforms state-of-the-art methods on all three benchmark datasets and stands out in recognizing minority emotions.

7. Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grants No.2021QY2102, National Natural Science Foundation of China under Grants No.62172089, No.61972087, Jiangsu Provincial Key Laboratory of Computer Networking Technology, Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No. BM2003201, Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No. 93K-9, and Nanjing Purple Mountain Laboratory.

8. Bibliographical References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. Openprompt: An open-source framework for prompt-learning. In *ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 105–113.

Paul Ekman. 1992. Are there basic emotions?

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowl. Based Syst.*, 248:108861.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP 2018, Brussels, Belgium, October 31 - November 4, 2018*, pages 2594–2604. 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 *NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7042–7052.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 397–406.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *NeurIPS 2020, December 6-12, 2020, virtual*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make BART a good dialogue emotion recognizer. In *AAAI 2022, February 22 - March 1, 2022*, pages 11002–11010.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995.

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *ACL 2022,* *Dublin, Ireland, May 22-27, 2022*, pages 5216–5228.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *CoRR*, abs/2103.10385.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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 *AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL 2021, Online, April 19 - 23, 2021*, pages 255–269.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *AAAI 2021, February 2-9, 2021*, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1551–1560.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5197–5206.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Huanqin Wu, Baijiaxin Ma, Wei Liu, Tao Chen, and Dan Nie. 2022. Fast and constrained absent keyphrase generation by prompt-based learning. In *AAAI 2022, February 22 - March 1, 2022*, pages 11495–11503.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2438–2447.

Yuting Yang, Wenqiang Lei, Pei Huang, Juan Cao, Jintao Li, and Tat-Seng Chua. 2023. A dual prompt learning framework for few-shot dialogue state tracking. In *WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1468–1477.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Jingjie Yi, Deqing Yang, Siyu Yuan, Kaiyan Cao, Zhiyao Zhang, and Yanghua Xiao. 2022. Contextual information and commonsense based prompt for emotion recognition in conversation. In *ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part II*, volume 13714 of *Lecture Notes in Computer Science*, pages 707–723.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421.