# Improving Contrastive Learning in Emotion Recognition in Conversation via Data Augmentation and Decoupled Neutral Emotion

**Yujin Kang, Yoon-Sik Cho**
Department of Artificial Intelligence
Chung-Ang University, South Korea
{zinzin32, yoonsik}@cau.ac.kr

## Abstract

Emotion recognition in conversation (ERC) has attracted much attention due to its wide applications. While consistent improvement is being made in this area, inevitable challenge comes from the dataset. The ERC dataset exhibits significantly imbalanced emotion distribution. While the utterances with *neutral* emotion predominate the data, this emotion label is always treated the same as other emotion labels in current approaches. To address the problem caused by the dataset, we propose a supervised contrastive learning specifically oriented for ERC task. We employ a novel data augmentation method emulating the emotion dynamics in a conversation and formulate supervised contrastive learning method tailored for ERC addressing the predominance and the ambiguity of neutral emotion. Experimental results on four benchmark datasets demonstrate the effectiveness of our approach.

## 1 Introduction

Motivated by the success of chatbot services, emotion recognition in conversation (ERC) has become an active research field, where the task is predicting the emotions in utterances in a conversation. The key point of ERC is how to effectively model the context of each utterance and corresponding speaker. In order to capture the contextual information, existing works generally resort to recurrence-based methods (Poria et al., 2017a; Majumder et al., 2019), graph-based methods (Ghosal et al., 2019; Shen et al., 2021b), knowledge-based methods (Zhong et al., 2019; Ghosal et al., 2020; Lee and Lee, 2022), and pre-trained language model (Kim and Vossen, 2021; Qin et al., 2023). Despite the improvements, there always remain intrinsic challenges in ERC dataset.

One challenge comes from the ERC dataset, where the emotion labels are often imbalanced. Previous studies (Yang et al., 2022; Gao et al., 2022) in ERC have pointed out that the imbalanced datasets cause negative impact on the prediction performance. Specifically, the class with the smallest number of samples suffers in the process due to the relatively insufficient amount of data for training. Some of the works (Guibon et al., 2021; Song et al., 2022a) have been introduced to overcome the limitation of the dataset. Guibon et al. (2021) use few-shot setting in episodic approach (Ravi and Larochelle, 2017), which simulates a context with only few examples per class; SPCL(Song et al., 2022a) leverages a prototype for each category as at least one positive sample of the same category and negative samples of all other categories in contrastive learning.

The second challenge of ERC comes from the emotion label *neutral* which is the majority class dominating the dataset yet indistinct. Prior studies have pointed out that the model tends to misclassify emotions to neutral (Majumder et al., 2019; Ghosal et al., 2019; Shen et al., 2021b). The main reason behind this is that the models tend to predict towards majority class, which is neutral in ERC. Besides, neutral was set as default emotion, where non-neutral emotions were annotated by human annotators only when the intensity of emotions (arousal) was sufficiently strong. While this setting can discern different emotions among non-neutral, the distinction between neutral and non-neutral becomes vague. Even with these challenges, most existing studies in ERC treat neutral emotions same as other non-neutral emotions for classification. Only a few recent works (Zhang et al., 2020a; Qin et al., 2023) have treated neutral in different ways from other emotions, such as alleviating the confusion between neutral and non-neutral through auxiliary tasks or detecting neutral first in coarse-grained level. However, due to the two-stage learning scheme, these models are inherently suboptimal.

To tackle the limitation of ERC dataset, we introduce a novel supervised Contrastive Learning

*Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 2194–2208
March 17-22, 2024 ©2024 Association for Computational Linguistics

framework which is specifically oriented for ERC Dataset (**CLED**). To address the first challenge of ERC dataset: data imbalance, CLED employs a novel data augmentation technique that utilizes the emotion centroids obtained from the pre-trained language model (PLM) embeddings. We perform interpolations on these centroids for generating augmented utterances, where the interpolation is performed reflecting the emotion shift through Markovian property. From the training data, each transition probability is computed, and is fully utilized for our data augmentation. Our method is unique in that the interpolation is based on the the realistic scenario with emotion shift.

We further address the second challenge of the ERC dataset concerning the limited use of the *neutral*. We design contrastive learning specifically dedicated to neutral emotion. We formulate an objective function that repels specific label, neutral, more strongly than others to clarify the boundaries of each label, considering that neutral closely intersects with other emotions. As such, CLED makes non-neutral emotions more distinct from neutral emotion by applying a stronger repelling force from neutral.

To verify the effectiveness, we implement our proposed scheme on six baselines including five recent ERC models and a RoBERTa-large based classifier which we additionally implement for this study. We compare the results using four benchmark ERC datasets. Experimental results show that two operations we propose consistently improves the performance. We further show that our method significantly outperforms other data augmentation method. Our contribution can be summarized into three-fold.

- We propose new contrastive learning for addressing the limitations of ERC dataset via data augmentation and decoupling neutral emotion from other emotions.
- To the best of our knowledge, this is the first attempt to apply data augmentation method for ERC. The augmentation is tailored for ERC reflecting the nature of conversation and the emotion shift.
- We conduct experiments with four benchmark datasets in ERC. Extensive experiments verify the effectiveness of our proposed method and demonstrate how each of the operations we introduce contributes to the model performance.

## 2 Related work

### 2.1 Emotion Recognition in Conversation

Existing ERC models resort to diverse deep learning method to effectively model dialogue, and can be devided into four groups: recurrent, graph, knowledge, and pre-trained language model(PLM) based methods.

Early studies consider utterances as sequential data through LSTM (Poria et al., 2017b) or the gated recurrence unit (GRU)-based model (Hazarika et al., 2018a,c; Majumder et al., 2019; Jiao et al., 2019). The graph-based methods (Ghosal et al., 2019; Zhang et al., 2019; Shen et al., 2021a) represent conversation as nodes and edges of a graph. Specifically, DAG-ERC (Shen et al., 2021b) models the conversation in directed acyclic graph, and combines recurrence and graph-based methods. The knowledge-enhanced methods leverage external knowledge by integrating it with hierarchical transformers (Zhong et al., 2019), capturing the complex interactions (Ghosal et al., 2020) and building structural psycological interactions (Li et al., 2021). Recent works use PLM as utterance encoder (Shen et al., 2021b; Li et al., 2021, 2022). Lee and Lee (2022) exploit PLM to model context and speaker's memory. Qin et al. (2023) integrate utterance, context, and dialogue structure information through fine-turning PLM.

While these studies have constantly made improvements in ERC, these models suffer from the class imbalance due to the predominance of neutral emotion. Some recent works try to mitigate these issues from the learning perspective. ProtoSeq (Guibon et al., 2021) adopts few-shot learning for resolving challenge. SPCL (Song et al., 2022a) tried to solve this problem by combining prototypical networks (Snell et al., 2017) with supervised contrastive learning (Khosla et al., 2020). However, these works more focus on non-neutral emotions which are relatively small compared to neutral emotion, where few-shot approaches are borrowed for handling few samples with non-neutral emotions. Besides, neutral emotion is treated in the same way as other non-neutral emotions.

### 2.2 Supervised Contrastive Learning

Contrastive learning brings an anchor and its augmented sample closer together, while simultaneously pushing the anchor away from negative samples in the embedding space. The supervised contrastive learning (SupCon) (Khosla et al., 2020) extends the self-supervised contrastive approach by considering data with the same label as positive samples and data with a different label than the anchor as negative samples.

Some researchers have implemented SupCon in the context of ERC. CoG-BART (Li et al., 2022) is the first attempt to apply SupCon to ERC to effectively identify similar emotions by mutually excluding different emotions. Song et al. (2022a) employ the supervised contrastive loss with a prototypical network to address imbalanced data integrating a curriculum learning strategy. Recently, Hu et al. (2023) propose supervised adversarial contrastive learning generating worst-case samples to ensure label-level consistency and fine-grained intra-class features.

### 2.3 Text Data Augmentation

Data augmentation can increase training data without directly collecting data (Feng et al., 2021). Data augmentation for text has been promoted with diverse approaches such as random word delection, swapping, insertion (Wei and Zou, 2019), back-translation (Sennrich et al., 2016), and erasing part of the information (Shen et al., 2020) to generate perturbed samples. Data augmentation approaches for conversational data have also been introduced and have effectively improve the performances. The existing works for conversational data augmentation mainly focus on task-oriented dialogue (Quan and Xiong, 2019), summarization (Chen and Yang, 2021), and dialogue generation (Hou et al., 2018; Zhang et al., 2020b). However, data augmentation for ERC, to the best of our knowledge, has not been previously studied. This can be challenging as emotion dynamics and context information are involved in a conversation.

Beside the word-level and sentence-level data augmentation, recent study (Chen et al., 2020) proposed a data augmentation approach through interpolation on hidden space. This idea is based on the manifold mixup (Verma et al., 2019) which improves generalization in deep neural networks.

## 3 Methodology

### 3.1 Problem Formulation

We assume ERC dataset is comprised of $\mathcal{D} = \{C_1, C_2, ..., C_{|\mathcal{D}|}\}$ which is a collection of $|\mathcal{D}|$ conversations. A conversation is a sequence of utterances $C = \{(u_1, s_1, y_1), (u_2, s_2, y_2), \ldots, (u_n, s_n, y_n)\}$, where $s_i$, $y_i$ represent the speaker and label of $u_i$ and $n$ denotes the number of utterances in conversation. Each utterance $u_i$ consists of sequence of tokens $u_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$, where $m$ is the number of tokens. When target utterance $(u_t, s_t)$ and its context $\{(u_1, s_1), (u_2, s_2) \ldots, (u_{t-1}, s_{t-1})\}$ is given, the goal of ERC is to predict emotion label $(y_t)$ of target $u_t$.

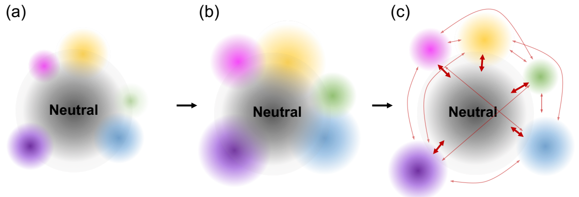


Figure 1: Intuition of our study for ERC dataset. The red lines represent pushing apart negative pairs, with the thickness of the lines indicating the intensity of the pushing force.

### 3.2 Overview

The entire process of our approach is depicted in Figure 1. Figure 1 is hypothetical visualization of the ERC data for illustration purposes, where each color represents different emotion labels. The data in Figure 1 (a) represents initial embedding, without undergoing any processing. In Figure 1 (b), the class imbalance problem is alleviated through proposed data augmentation scheme in our CLED. However, the datapoints with *neutral* is hardly differentiated due to the nature of ERC dataset. Our proposed contrastive learning approach specifically applies stronger repulsion force to attack this problem as shown in Figure 1 (c). In the following, we provide further details of each operation.

### 3.3 Data Augmentation for ERC

Here we introduce a data augmentation method tailored to ERC. Our data augmentation, inspired by TMix (Chen et al., 2020), augments data in hidden space through interpolation. Unlike TMix, which generates data from independent sentences, we perform sequence-level data augmentation. As utterances are processed as a sequence for ERC, we leverage the hidden space representation for contextual modeling and capturing emotion dynamics.

Specifically, our approach emulates how emotions are induced in a conversation. The emotion of the next utterance is affected by the current utterance. Our method generates virtual training samples through linear interpolations with each cen-

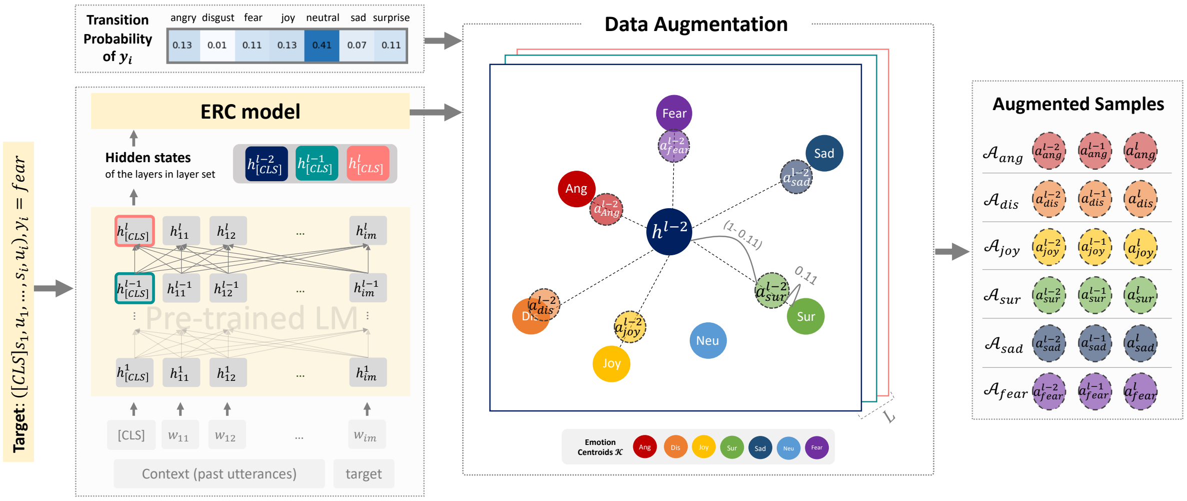


Figure 2: Framework of our data augmentation method. For each non-neutral emotion class, augmented embeddings can be obtained. $l$ represents the number of layers in the PLM. We consider the layer set $L$ as $\{l-2, l-1, l\}$ in this figure. As shown in the right-most figure, we perform data augmentation for each layer in the layer set.

troid and the current utterance embedding, which can be viewed as generating new utterance for next step from the current context ($h_i$). The overall process for our data augmentation is summarized in Figure 2, where the centroid for each emotions and transition matrix are used for interpolations.

**Data Augmentation on hidden space** We bring pre-trained language model (PLM) as an embedding module. We use the special token [CLS] to get embeddings that reflect context information. In the embedding stage, the input and output of $u_i$ are as follows:

$$\text{PLM}([\text{CLS}], I_{\text{ERC}}) = \{h_i^l | l \in L\}, \qquad (1)$$

where $I_{\text{ERC}}$ is the model-specific input and $h_i^l$ is the embedding of [CLS] of $u_i$ in $l$-th hidden layer. $L$ is the layer set we select to make the hidden state of the target for augmentation. Since the PLM is a multi-layer model, we can obtain diverse embeddings from each hidden layer for the same input. In Section 5.7, we elaborate how multiple combinations of hidden layers for $h_i$ have been tried, and report each performance.

**Emotion centroids** We generate augmented data around emotion centroids. We take inspirations from prototypical networks (Snell et al., 2017; Guibon et al., 2021), and borrow the idea of prototype. For each class of emotions, we collect all utterance embeddings associated to given emotion label, and compute the centroid for each. The set of centroids can be expressed as follows:

$$\mathcal{K} = \{\frac{1}{|\{(h_i, y_i)|y_i = e\}|} \sum_{(h_i, y_i), y_i = e} h_i | e \in \mathcal{E}\}, \qquad (2)$$

where $\mathcal{K}$ is the set of emotion centroids. $\mathcal{E}$ is the emotion label set.

**Interpolations with emotion shift** Prior works have found that the emotions in dialogue have a dependency: the inter- and intra-speaker dependency (Hazarika et al., 2018b; Wang et al., 2020; Ghosal et al., 2021) and label copying property (Navarretta, 2016; Poria et al., 2019b; Song et al., 2022b). The $u_1$'s emotion affects $u_2$, and this process sequentially continues throughout the conversation. Based on this, we represent sequential emotion dependency as a Markovian transition matrix. We count the current emotion changes to each subsequent emotion in the dialogue and convert them to probability from the training data. The transition matrix illustrates how the current emotion $i$ changes to subsequent emotion $j$ in the dialogue. The detailed information about the transition matrix can be founded in Appendix A.

With the computed transition matrix, we perform interpolation between each of the emotion centroids and the $h_i$, embedding of $u_i$. Given embedding $h_i$, virtual sample with emotion label $j$ is augmented as below.

$$a_{j|i} = \lambda_{ij} h_i + (1 - \lambda_{ij}) k_j, \qquad (3)$$

where $\lambda_{ij}$ represents the value corresponding emotion $j$ of the row of $y_i$ in the transition matrix and

**Algorithm 1** Learning procedure at each epoch. Note that the Encoder($\cdot$) can be of any type.

```
Input:
Training dataset $\mathcal{D}$ = $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{|\mathcal{D}|}\}$ ;
$\mathcal{T}$: Transition matrix;
$\mathcal{E}$: label set;
Encoder: ERC model
Output: A prediction $P$ for utterances in conversation
 1: $\mathcal{A}$ = []
 2: calculate emotion centroids $\mathcal{K}$ (Eq 2)
 3: $\mathcal{D}_{\text{PLM}} \leftarrow$ PLM($\mathcal{D}$) (Eq 1)
 4: $\mathcal{D}_{\text{ebd}} \leftarrow$ Encoder($\mathcal{D}_{\text{PLM}}$)
 5: for $(\mathcal{H}_i, y_i) \in \mathcal{D}_{\text{ebd}}$ do
 6:   for $j \in$ ($\mathcal{E}$- neutral) do
 7:     $\lambda_{ij}$ = $\mathcal{T}[y_i][j]$
 8:     for $h_i^l \in \mathcal{H}_i$ do
 9:       $a_{j|i}$ = $\lambda_{ij}$ $h_i^l$ + (1-$\lambda_{ij}$) $k_j$ (Eq 3)
10:       $\mathcal{A}$.append($a_{j|i}$)
11:     end for
12:   end for
13: end for
14: $\mathcal{L}_{\text{Encoder}}$ = Encoder loss($\mathcal{D}_{\text{ebd}}$)
15: $\mathcal{L}_{\text{CLED}}$ = CLED loss($\mathcal{D}_{\text{ebd}}$, $\mathcal{A}$) (Eq 4-9)
16: $\mathcal{L}$ = $\mathcal{L}_{\text{Encoder}}$ + $\mathcal{L}_{\text{CLED}}$
17: Optimize PLM and Enocder
```

$k_j \in \mathcal{K}$ is the centroidal emotion $j$. If $\lambda$ is large, new data with emotion $j$ may be greatly affected by $h_i$. Since our strategy uses interpolating points presenting emotion dependency in conversation and yields augmentation by embedding of PLM, it renders generated samples specific for ERC.

For a given utterance, the number of interpolated embeddings is determined by the product of the number of hidden layers and the count of emotion labels. We exclude neutral for this augmentation not to cause more data imbalance. The generated data from above scheme is later used along with the original data for calculating the contrastive loss during model training. Our CLED framework is detailed in the Algorithm 1, and the process of generating virtual data is outlined in Lines 5–13.

### 3.4 CLED: Supervised Contrastive Learning for ERC Dataset

While neutral emotion is assigned as default in ERC, many existing studies treat it the same as other non-neutral emotions in their training processes. Some studies simply exclude neutral from their evaluation. In this study, we attack the problem from different point of view, which is motivated by two observations. We observe that human annotators assigned neutral label when the utterance exhibits weak emotions (Kleinsmith et al., 2005) or when the utterance cannot be assigned to any of the candidates in non-neutral emotions (Zahiri and Choi, 2017; Li et al., 2017). In other words, neutral utterances are challenging to discern and even hinders learning other non-neutral emotions. To alleviate this problem, CLED reformulates supervised contrastive learning to concentrate on decoupling neutral emotion from other non-neutral emotions.

The supervised contrastive learning (SupCon) (Khosla et al., 2020) pulls anchor and samples with same label, and pushes samples with different label from anchor. Given sample $h_i$, which is the embedding of $u_i$, SupCon calculates positive and negative scores for contrastive loss as follows.

$$\mathcal{F}(h_i, h_j) = \exp(\cos(h_i, h_j)/\tau). \tag{4}$$

$\mathcal{F}$ is computed using a cosine similarity with temperature $\tau$ between two instances.

$$\mathcal{P}(i) = \sum_{h_p \in P(i)} \mathcal{F}(h_i, h_p). \tag{5}$$

$$\mathcal{N}_{sup}(i) = \sum_{h_j \in A(i)} \mathcal{F}(h_i, h_j). \tag{6}$$

$P(i)$ in Equation 5 is the set of positive samples with same labels with $h_i$ including virtual data generated by our data augmentation. In Equation 6, A(i) represents the negative set, comprising samples and augmented data with different labels from $h_i$. In Equation 5 and 6, both scores are sum of similarity between the anchor and samples.

The *neutral* within the ERC dataset shares some degree of similarity with all other emotions (Yang et al., 2022), and overlaps relatively with other emotions in the embedding space (Joshi et al., 2022). If all data in negative set is pushed by the same force regardless of label, the space between the non-neutral labels can be relatively easily separated. However, the neutral data still share the area with other labels. As shown in Appendix B.1, attempting to repel negative pairs by merely adjusting the hyperparameter $\tau$ leads to worse performance than SupCon. To comprehend the nature of neutral and effectively segregate it from other emotional areas, we have introduced an additional negative score tailored specifically for *neutral*. Based on SupCon, we tweak the Equation 6 to calculate neutral score.

| Dataset | Conversations |  |  | Utterances |  |  | $\lvert\mathcal{E}\rvert$ | Neutral | Imbalance | Evaluation |
|---|---|---|---|---|---|---|---|---|---|---|
|  | train | val | test | train | val | test |  | (%) | ratio | Metric |
| IEMOCAP | 108 | 12 | 31 | 5163 | 647 | 1623 | 6 | 22.98 | 3:1 | Weighted-F1 |
| EmoryNLP | 713 | 99 | 85 | 9934 | 1344 | 1328 | 7 | 29.95 | 4:1 | Weighted-F1 |
| MELD | 1038 | 114 | 280 | 9989 | 1109 | 2610 | 7 | 48.21 | 18:1 | Weighted-F1 |
| DailyDialog | 11118 | 1000 | 1000 | 87170 | 8069 | 7740 | 7 | 83.24 | 1156:1 | Micro-F1 |

Table 1: Statistics of ERC datasets. Neutral denotes the percentage of utterances with label-neutral. Imbalance ratio indicates the maximum class imbalance and cls is the number of different labels in each dataset.

$$\mathcal{N}_{neu}(i) = \sum_{h_j \in \{(h_j, y_j) | y_j = \text{Neutral}\}} \mathcal{F}(h_i, h_j). \quad (7)$$

$$\mathcal{N}(i) = \mathcal{N}_{sup}(i) + \alpha \mathcal{N}_{neu}(i). \quad (8)$$

Equation 7 collects similarity scores between $h_i$ and samples with the neutral label to repel non-emotional label from other emotions. In Equation 8, we combine the *neutral score* with the negative score of SupCon, which will bring additional repelling force specifically for neutral emotion. We adjust the force for neutral through parameter $\alpha$. Finally, the loss we optimize is presented below.

$$\mathcal{L}_{\text{CLED}}(i) = -\log\left(\frac{1}{|P(i)|} \cdot \frac{\mathcal{P}(i)}{\mathcal{N}(i)}\right). \quad (9)$$

$$\mathcal{L}(i) = \mathcal{L}_{\text{Encoder}}(i) + \mathcal{L}_{\text{CLED}}(i), \quad (10)$$

where $\mathcal{L}_{\text{Encoder}}$ represents the loss of the ERC model, which is typically a cross-entropy loss. We integrate our CLED loss with the ERC objective to effectively address the challenges posed by ERC dataset.

# 4 Experimental Settings

## 4.1 Datasets

We conduct experiments on four ERC benchmark dataset: IEMOCAP, EmoryNLP, MELD and DailyDialog. Table 1 shows the statistics of each dataset. For evaluation metrics, following previous studies, we employ micro-F1 excluding the majority class (neutral) for DailyDialog and weighted-F1 for other ERC datasets.

**IEMOCAP** (Busso et al., 2008) is a dyadic multimodal dataset that contains text, audio, video, and motion capture information movement. For ERC task, we bring only text data. The label set contains *happy, sad, angry, excited, frustrated*, and *neutral*.

**EmoryNLP** (Zahiri and Choi, 2017) is a text dataset extracted from the TV show *Friends* transcripts. Each utterance is labelled with *sad, scared, mad, powerful, peaceful, joyful*, and *neutral*. Their annotation is based on Willcox's feeling wheel (Willcox, 1982).

**MELD** (Poria et al., 2019a) is a multi-party multimodal dataset collected from the popular TV-series *Friends*. Each utterance has been annotated with one of *anger, disgust, fear, joy, surprise, sadness*, and *neutral*.

**DailyDialog** (Li et al., 2017) is a dyadic text dataset. The label set contains *anger, disgust, fear, joy, surprise, sadness*, and *neutral*, which are from the six Ekman's basic emotions (Ekman et al., 1999) and others.

## 4.2 Baselines

We apply our learning scheme to strong baseline models which is summarized here. For strict comparison, we use the exact hyper-parameter values employed by the original models, and do not perform any further tuning of each model when applying our scheme.

**RoBERTa** (Liu et al., 2019) is a pre-trained language model(PLM). We leverage RoBERTa-large [1] as the embedding module. [2] Classification layer is mounted to PLM for predicting emotion labels.

**Psychological** (Li et al., 2021) proposes a psychological-knowledge-aware interaction graph enhanced by commonsense knowledge and graph transformer.

**CoMPM** (Lee and Lee, 2022) uses the pre-trained memory as external knowledge utilizing the PLM as an extractor, which is combined with context model for ERC.

**EmoNotOne-SA** (Lee, 2022) tries to represent emotion as grayscale label and introduces several strategies for constructing grayscale label. We choose the self-adjust-grayscale method, which performs best among the grayscale construction

[1] https://huggingface.co/roberta-large

[2] We make the input by prepending the speaker for each utterance and concatenating a context, previous utterances of the target, to the current utterance.

| Model | Dataset | | | |
|---|---|---|---|---|
| | IEMOCAP | EmoryNLP | MELD | DailyDialog |
| RoBERTa (Liu et al., 2019) | 61.92 | 34.62 | 64.52 | 60.36 |
| Psychological (Li et al., 2021) | 63.03 | 37.44 | 63.52 | 58.93 |
| CoMPM (Lee and Lee, 2022) | 66.33 | 37.06 | 65.19 | 59.01 |
| EmoNotOne-SA (Lee, 2022) | 62.51 | 36.38 | 65.00 | 60.71 |
| EmotionFlow (Song et al., 2022b) | - | 38.61 | 66.00 | - |
| SPCL (Song et al., 2022a) | 67.30 | 39.89 | 66.16 | - |
| RoBERTa+CLED | 62.77 (↑ 0.85) | 36.89 (↑ 2.27) | 66.24 (↑ 1.72) | 61.23 (↑ 0.87) |
| Psychological+CLED | 64.03 (↑ 1.00) | 37.90 (↑ 0.46) | 64.09 (↑ 0.57) | 59.49 (↑ 0.56) |
| CoMPM+CLED | **67.65** (↑ 1.32) | 38.76 (↑ 1.70) | 66.00 (↑ 0.81) | **61.57** (↑ 2.56) |
| EmoNotOne-SA+CLED | 63.63 (↑ 1.12) | 37.71 (↑ 1.33) | 65.61 (↑ 0.61) | 60.98( ↑ 0.27) |
| EmotionFlow+CLED | - | 40.54 (↑ 1.93) | **66.77** (↑ 0.77) | - |
| SPCL+CLED | 66.58 (↓ 0.72) | **40.76** (↑ 0.87) | 66.44 (↑ 0.28) | - |

Table 2: The results on ERC models on four benchmark datasets. We result in score of the average of five runs. Bold score indicates the best performance in each dataset. All of the results in baselines are reproduced by our experimentation with the original code.

| Method | IEMOCAP | EmoryNLP | MELD | DailyDialog |
|---|---|---|---|---|
| CLED | 62.77 | 36.89 | 66.24 | 61.23 |
| - Data Augmentation (DA) | 62.30 (↓ 0.47) | 34.51 (↓ 2.38) | 65.31 (↓ 0.93) | 59.67 (↓ 1.56) |
| - Neutral Score (NS) | 61.87 (↓ 0.90) | 35.74 (↓ 1.15) | 64.82 (↓ 1.42) | 60.56 (↓ 0.47) |
| - DA - NS | 61.58 (↓ 1.19) | 34.58 (↓ 2.31) | 64.14 (↓ 2.10) | 57.46 (↓ 3.77) |

Table 3: Ablation study. The numbers in parentheses indicate the difference in performance when the component is removed from CLED during training.

methods.

**SPCL** (Song et al., 2022a) employs the supervised contrastive learning loss combining prototypical network and curriculum learning for mitigating data imbalance and handling few samples in ERC dataset.

**EmotionFlow** (Song et al., 2022b) is a model considering the spread of speakers' emotions, and further utilizes the Conditional Random Field (CRF) to capture sequential emotional information.

### 4.3 Implementation Details

When we implement the RoBERTa-large, we set the learning rate to 1e-6. The number of epochs and batch-sizes are 10 and 8, respectively. Otherwise, we follow the original setting of baseline models. We train and test the model on a single Nvidia A100. We fix $\tau$ in Equation 4 to 0.05. For $\alpha$ in Equation 8, we search the parameter using the validation set. In general, fixing $\alpha$ to larger value with respect to the percentage of neutral leads to better performance. In our experiments, $\alpha$ is set to 0.9 for DailyDialog, where about 83% of the data are tagged as neutral; 0.2 for EmoryNLP which has a relatively small ratio of neutral. It is worth noting that $\alpha$ can be translated as the *additional* force on neutral, and alpha has (1+$\alpha$) effect overall. In Section 5.6, we show how CLED is robust to various settings of $\alpha$.

## 5 Experiments

### 5.1 Comparisons with State-of-the-art Methods

Our proposed method is model-agnostic, where we can apply to existing approaches in ERC. We use six baselines and compare each of performance implementing our approach as plug-and-play to the original baselines. These baseline models are selected from the ERC literature which have achieved state-of-the-art results and leveraged PLM in their embedding methods. The models without released code are not included in the experiment. Table 2 shows the efficacy of our approach, where we constantly achieve performance improvements on all the baselines except one model on one of the dataset. The best performing results for each dataset is reported in bold. It is also worth not-

| Method | Dataset | | | |
|---|---|---|---|---|
| | IEMOCAP | EmoryNLP | MELD | DailyDialog |
| RoBERTa (base) | 61.58 | 34.58 | 64.14 | 57.46 |
| Random Delete | 59.72 | 27.55 | 64.41 | 60.16 |
| Random Swap | 58.87 | 32.33 | 64.51 | 59.74 |
| Random Insert | 61.68 | 33.39 | 64.18 | 59.21 |
| Synonym Replacement | 60.14 | 32.38 | 64.42 | 59.94 |
| Dropout | 60.50 | 31.09 | 64.75 | 59.52 |
| Our DA | **61.87** | **35.74** | **64.82** | **60.56** |

Table 4: The comparison with other data augmentation methods.

ing that even the simple RoBERTa model with our approach already outstands some competitive baselines. However, when we apply our scheme on SPCL, the performance drops from the original model. The reason may be that the IEMOCAP dataset is relatively balanced, and SPCL was achieving best performance among the other baselines for IEMOCAP benefiting from other aspects leaving only little room for improvement through our scheme.

### 5.2 Ablation Study

Here we conduct an ablation study to provide the empirical evidence of our claim by stripping each component. We apply our method to RoBERTa using all benchmark datasets. Table 3 illustrates the contribution of each operation in CLED to the model's performance. When we exclude samples generated through data augmentation from CLED, the performance is lower across all benchmarks. Not combining the *neutral score* in Equation 7 with the negative score in SupCon results in a consistent performance decline across all datasets. Furthermore, removing our two components (i.e., RoBERTa with vanilla-supervised contrastive learning) leads to inferior performance, highlighting the effectiveness of both data augmentation and the decoupling of neutral.

### 5.3 Comparison with Different Types of Data Augmentation

To better demonstrate the effectiveness of our data augmentation approach, we further compare it with other data augmentation methods. We use the RoBERTa with SupCon loss as base model for this experiment. Table 4 shows the performances with diverse data augmentation methods. Random delete, swap, insert, and synonym replacement from EDA (Wei and Zou, 2019) are the techniques that randomly choose $n$ words from the conversation and transform them. We believe that context understanding and emotional dependency are the core components of conversation, which is also confirmed in the results in Table 4. We additionally test dropout [3] which has been proposed in SimCSE (Gao et al., 2021). Our data augmentation outperforms Dropout on all datasets.

| Emotion | RoBERTa | + CLED |
|---|---|---|
| fear (1.92%) | 11.76 | 23.38 (↑ 11.62) |
| disgust (2.61%) | 21.69 | 21.98 (↑ 0.29) |
| sadness (7.97%) | 44.86 | 45.34 (↑ 0.48) |
| surprise (10.77%) | 60.01 | 61.94 (↑ 1.93) |
| angry (13.22%) | 49.92 | 53.35 (↑ 3.43) |
| joy (15.4%) | 61.48 | 62.67 (↑ 1.19) |
| neutral (48.12%) | 78.05 | 79.45 (↑ 1.40) |

Table 5: Comparison of performance with and without our method in MELD dataset. A parenthesis next to each emotion label represents the percentage of emotion label within test set.

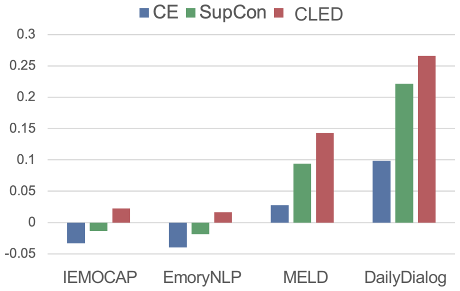


Figure 3: The silhouette score for each of three losses.

### 5.4 Performance on Minor Label

We report how our proposed model performs on each class label from MELD dataset, where our scheme always improve the performance on every emotion classes. As presented in Table 5, the result of RoBERTa shows that classifying minor labels is challenging. Compared with the performance of the major label (neutral), the performance of the minor label (fear) decreases dramatically by 66.29%.

When we combine RoBERTa with our method, the model is consistently superior to the vanilla-RoBERTa for all emotions. Specifically, our method shows significant performance improvement on a label with the least occurrence, fear, which is from 11.76% to 23.38%. We infer that the model attains a performance boost through more samples by augmentation. These augmented data help the model to classify unfamiliar labels.

[3] The SimCSE tries out different dropout rates and finds that dropout probability $p = 0.1$ performs best. Following this, we use a dropout rate of 0.1

| $\alpha$ | Dataset | |
|---|---|---|
| | MELD | EmoryNLP |
| 0 (base) | 64.82 | 35.74 |
| + additional repelling | | |
| 0.2 | 65.50 | **36.89** |
| 0.4 | 65.99 | 36.74 |
| 0.6 | 65.66 | 36.65 |
| 0.8 | **66.24** | 36.26 |
| 1.0 | 65.69 | 36.10 |

Table 6: Performances across different hyperparameter values on MELD and EmoryNLP. We report the weighted-F1 score.

### 5.5 Silhouette Score on Neutral Label

According to Yang et al. (2022), the neutral is similar to other label to some extent and is spread in overlapped with others. We compute the silhouette scores (Rousseeuw, 1987) on embeddings with neutral emotion to numerically validate the effectiveness of CLED, which is presented in Figure 3.

We compare the scores from RoBERTa optimized with three different losses: cross entropy(CE), SupCon, and CLED. Figure 3 shows that the silhouette score of our CLED outperforms the scores obtained through other losses on all datasets. As CE does not act against neutral, the neutral is spread on the embedding space, leading to the lowest score among the three objective functions. Compared to SupCon, CLED concentrate on repelling neutral from other emotions. Additionally, we visualize the representation with CE and CLED to qualitatively evaluate our loss in Appendix B.2.

### 5.6 Sensitivity Analysis on parameter $\alpha$

In Table 6, we perform a sensitivity analysis on the parameter ($\alpha$) that controls the imposition of extra negative scores for neutral instances in Equation 8. We conduct experiment using the two representative datasets: MELD dataset, which exhibits a data distribution significantly skewed to a neutral label, and the EmoryNLP dataset, which has a comparatively more balanced label distribution.

When $\alpha$ is set to 0, the loss becomes equivalent to the conventional contrastive learning setting. Setting $\alpha$ higher than 0 means infusing additional repeling force around neutral label. We achieve consistent performance improvement in every setting of $\alpha$, which can reflect the robustness of CLED. We believe these results are meaningful in that we always achieve performance improvement even across different datasets. We also highlight that the optimal $\alpha$ from different datasets indirectly reflects data characteristics in terms of neutral proportion. The optimal $\alpha$ in MELD is 0.8 and the optimal $\alpha$ in EmoryNLP is 0.2.

| Layer set | # layers | F1-score |
|---|---|---|
| {23} (last layer) | 1 | 64.82 |
| {5,6,7} | 3 | **65.97** |
| {19,20,21} | 3 | 65.25 |
| {0 - 11} | 12 | 65.38 |
| {12 - 23} | 12 | 65.76 |

Table 7: Comparison of performance with diverse layer combination in MELD dataset.

### 5.7 Comparison for Layer Set in Data Augmentation

Throughout all the experiments above, we only took the last layer of the PLM for data augmentation. Our augmentation can be further improved by finding the optimal combinations of hidden layer, which could be dependent on PLM or dataset. In our main results, we didn't search for the best combinations. Here we follow the study in (Jawahar et al., 2019), and investigate different configurations for layer set $L$ as shown in Table 7. We use the RoBERTa-large model, including a data augmentation component, as our base model to verify the effect of the number of layers. The results in Table 7 suggest possible directions for future work. Our model achieves the best performance with $L = \{5, 6, 7\}$. More details can be found in Appendix B.3.

## 6 Conclusion

In this paper, we discuss the challenges in the ERC dataset, which exhibits imbalanced label distribution and a dominance of neutral emotions that are difficult to distinguish from other emotions. We introduce novel method CLED to address the challenges of ERC datasets. The CLED employs a novel data augmentation reflecting the context and emotion-dependency in conversation. With augmented data, we redefine a supervised contrastive learning loss specifically designed for the ERC dataset to better distinguish between neutral and non-neutral emotions. We conduct extensive experiments to verify the effectiveness of our approach by constantly improving the previous baselines through plug-and-play.

## Limitations

This study has two limitations. 1) As our proposed data augmentation method is based on the pre-trained model, it can only combine with the model that leverages the pre-trained language model as their embedding module, and human evaluation on data augmentation is not available. 2) While performing data augmentation on the last layer of PLM is sufficiently effective, we verify that more layers boost the performance in Section 5.7. However, leveraging more hidden states of PLM increases computational resources. The tradeoff between performance and computational cost should be considered.

## Acknowledgements

This work was supported in by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang University)) and (No. 2022-0-00331, Development of Emotion Recognition/Generation-Based Interacting Edge Device Technology for Mental Health Care)

## References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

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. *Language resources and evaluation*, 42:335–359.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

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. *Knowledge-Based Systems*, 248:108861.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL).

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449.

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

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6858–6870.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018b. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018c. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. 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)*, pages 397–406.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Andrea Kleinsmith, P Ravindra De Silva, and Nadia Bianchi-Berthouze. 2005. Grounding affective dimensions into posture features. In *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*, pages 263–270. Springer.

Joosung Lee. 2022. The Emotion is Not One-hot Encoding: Learning with Grayscale Label for Emotion Recognition in Conversation. In *Proc. Interspeech 2022*, pages 141–145.

Joosung Lee and Wooin Lee. 2022. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 1204–1214.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11002–11010.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 469–474.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

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

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

Xiangyu Qin, Zhiyu Wu, Jinshi Cui, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, and Li Wang. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. *arXiv preprint arXiv:2301.06745*.

Jun Quan and Deyi Xiong. 2019. Effective data augmentation approaches to end-to-end task-oriented dialogue. In *2019 International Conference on Asian Language Processing (IALP)*, pages 47–52.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *International conference on learning representations*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

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

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

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

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022b. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546. IEEE.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Gloria Willcox. 1982. The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11595–11603.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

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*, pages 5415–5421.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020a. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020b. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

# Appendix

## A Transition Matrix in ERC dataset

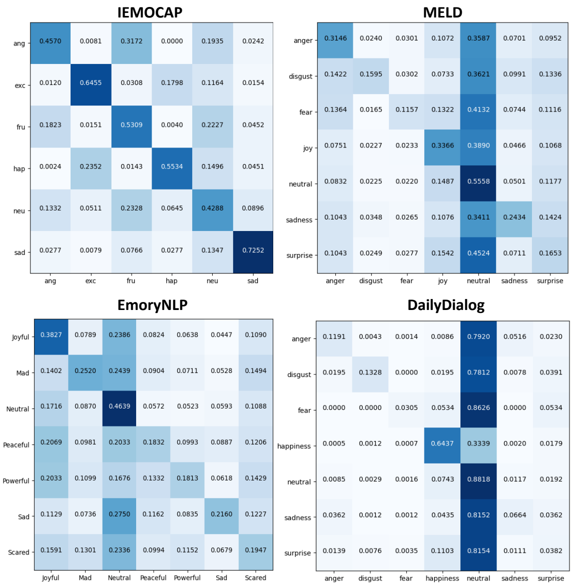


Figure 4: Transition Matrix

The transition matrix illustrates how the current emotion in row $i$ changes to subsequent emotion in column $j$ in the dialogue. Figure 4 represents the transition matrices calculated from the training data of each ERC dataset. The transition matrix's rows represent the probabilities of moving from one emotion to another, with each row summing to 1. For example, in the transition matrix of the IEMOCAP dataset, the first row indicates the probabilities of the next utterance's emotions given the current emotion is anger. The probability value for transitioning from anger to frustrated is 0.31.

## B Extended Experiments

Here we provide more experiments to prove effectiveness of our CLED.

### B.1 Effectiveness of CLED over Other CL Approach with Strong Negative

| Loss | *w/o* DA | *with* DA |
|---|---|---|
| SupCon | 64.14 | 64.82 |
| All-CL | 63.58 (-0.56) | 65.64 (+0.82) |
| CLED | 65.31 (+1.17) | **66.24** (+1.42) |

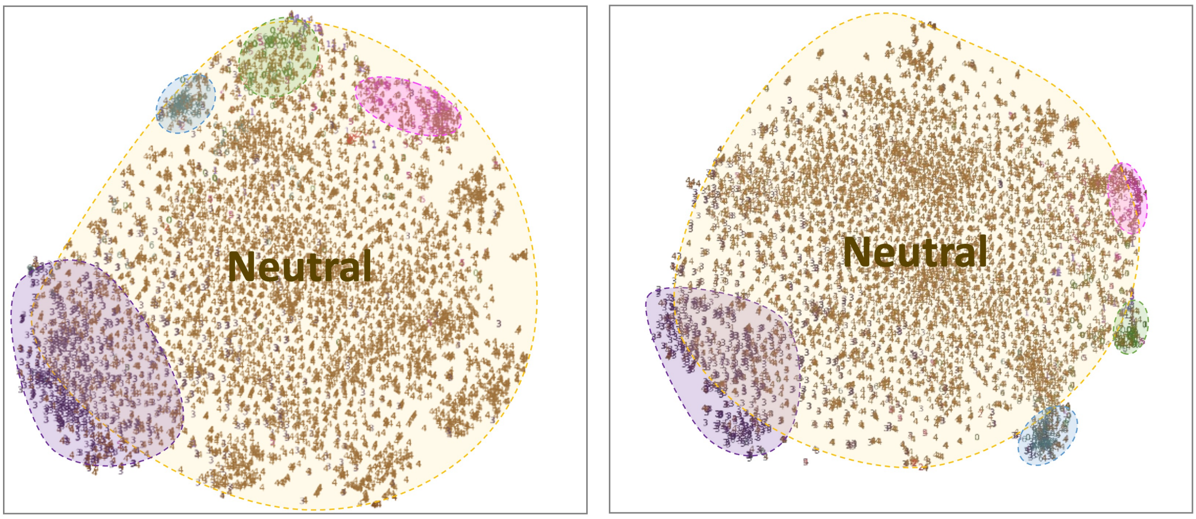
Table 8: Comparison of performance in MELD dataset. The parenthes indicate the difference in performance with Supcon.

Table 8 compares the performance with three-loss strategies: vanilla-supervised contrastive learning(SupCon), supervised contrastive learning applying more weights across all negative paris (All-CL), and CLED, applying only weight for the neutral. We use $\alpha$ by 0.8, the parameter to control the force, to All-CL and CLED.

When All-CL trains the model without data augmentation, the performance is degraded more than SupCon. Although the All-CL with data augmentation improves the performance, our CLED outperforms All-CL. This result indicates that repelling the negative pairs more strongly without considering the label property worsens the performance. Since neutral is nearly contacted with other emotions, our tailored contrastive learning method which pinpoints to neutral is more effective.

### B.2 Representation Visualization

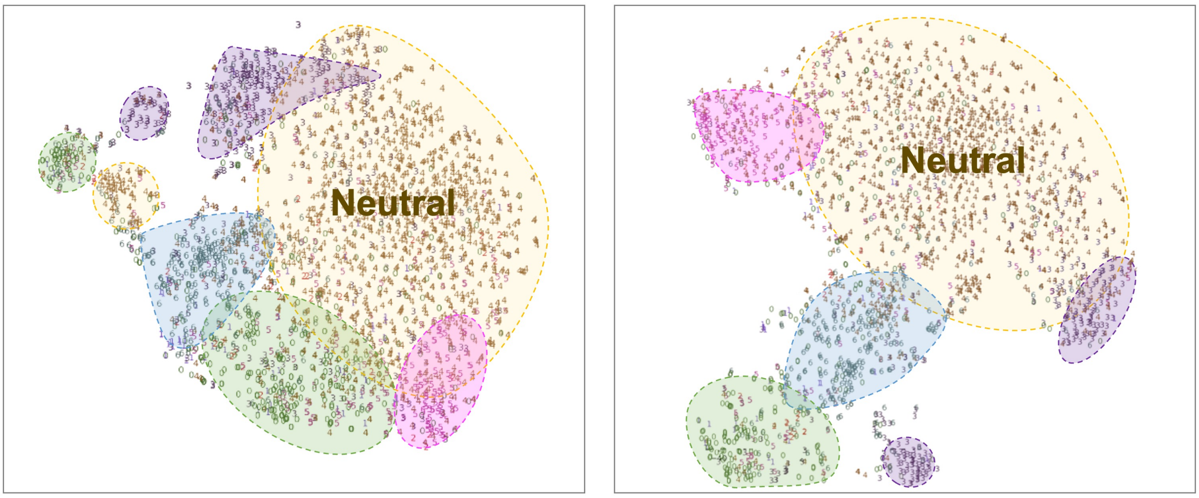
We visualize the learned representations with t-SNE (Van der Maaten and Hinton, 2008) on the test set of MELD and DailyDialog, where the neutral label accounts for more than a half of data. To clarify effectiveness of our loss tailored to neutral, we compare two variants: RoBERTa trained by cross-entropy(CE) and CLED. The results are shown in Figure 5 and 6. When the model is trained by CE, we observe that the neutral label is spread in the embedding space and overlaps with other labels. Our CLED makes the neutral relatively tight and united than CE loss. Thus, we can obtain the more apparent boundary of each emotion shown in (b) of Figure 5 and 6.



(a) CE

(b) CLED

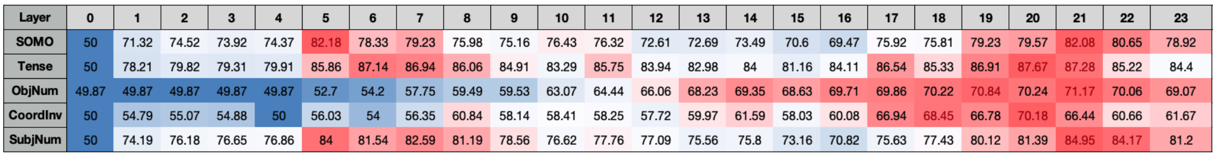
Figure 5: The representations learned with different optimization objectives on DailyDialog dataset. Each color represents different emotion label; The marking indicates that each label dominates this space.



(a) CE

(b) CLED

Figure 6: The representations learned with different optimization objectives on MELD dataset. Each color represents different emotion label. The marking indicates that each label dominates this space.



| Layer | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SOMO | 50 | 71.32 | 74.52 | 73.92 | 74.37 | 82.18 | 78.33 | 79.23 | 75.98 | 75.16 | 76.43 | 76.32 | 72.61 | 72.69 | 73.49 | 70.6 | 69.47 | 75.92 | 75.81 | 79.23 | 79.57 | 82.08 | 80.65 | 78.92 |
| Tense | 50 | 78.21 | 79.82 | 79.31 | 79.91 | 85.86 | 87.14 | 86.94 | 86.06 | 84.91 | 83.29 | 85.75 | 83.94 | 82.98 | 84 | 81.16 | 84.11 | 86.54 | 85.33 | 86.91 | 87.67 | 87.28 | 85.22 | 84.4 |
| ObjNum | 49.87 | 49.87 | 49.87 | 49.87 | 49.87 | 52.7 | 54.2 | 57.75 | 59.49 | 59.53 | 63.07 | 64.44 | 66.06 | 68.23 | 69.35 | 68.63 | 69.71 | 69.86 | 70.22 | 70.84 | 70.24 | 71.17 | 70.06 | 69.07 |
| CoordInv | 50 | 54.79 | 55.07 | 54.88 | 50 | 56.03 | 54 | 56.35 | 60.84 | 58.14 | 58.41 | 58.25 | 57.72 | 59.97 | 61.59 | 58.03 | 60.08 | 66.94 | 68.45 | 66.78 | 70.18 | 66.44 | 60.66 | 61.67 |
| SubjNum | 50 | 74.19 | 76.18 | 76.65 | 76.86 | 84 | 81.54 | 82.59 | 81.19 | 78.56 | 76.62 | 77.76 | 77.09 | 75.56 | 75.8 | 73.16 | 70.82 | 75.63 | 77.43 | 80.12 | 81.39 | 84.95 | 84.17 | 81.2 |

Figure 7: The result of probing for RoBERTa-large model.

### B.3 Probing hidden layer of RoBERTa

Probing tasks unearth the linguistic features possibly encoded in neural models (Adi et al.; Conneau et al., 2018; Jawahar et al., 2019). Jawahar et al. (2019) try to explain what linguistic features the intermediate layer of BERT contains and find that the bottom, middle, and top layers in BERT contain surface, syntactic, and semantic features, respectively. In order to find effective layer for augmentation, we bring Jawahar et al. (2019)'s method and search which layer in RoBERTa-large's intermediate layers is meaningful. The number of RoBERTa-large model's layers is 24.

As conversation is a sequence of utterances, we perform semantic probing. The detail information for each probing is as follows:

- **Tense**
- **SubjNum and ObjNum**: the subject and the object number in the main clause
- **SOMO**: the sensitivity to random replacement of a noun/verb
- **CoordInv**: the random swapping of coordinated clausal conjuncts.

Figure 7 shows the results of the probing task and indicates that {5, 6, 7, 19, 20, 21, 22, 23} layers contain semantic information. Although {5,6,7} layers do not predict well in ObjNum and Coordinv, the results in Table 7 show that Tense, SubjNum, and CoordInv are more important to the conversation.