Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations

Dou Hu[1,2] and **Yinan Bao**[1,2] and **Lingwei Wei**[1,2] and **Wei Zhou**[1*] and **Songlin Hu**[1,2*]

[1] Institute of Information Engineering, Chinese Academy of Sciences

[2] School of Cyber Security, University of Chinese Academy of Sciences

{hudou, baoyinan, weilingwei, zhouwei, husonglin}@iie.ac.cn

Abstract

Extracting generalized and robust representations is a major challenge in emotion recognition in conversations (ERC). To address this, we propose a supervised adversarial contrastive learning (SACL) framework for learning class-spread structured representations in a supervised manner. SACL applies contrast-aware adversarial training to generate worst-case samples and uses joint class-spread contrastive learning to extract structured representations. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features. To avoid the negative impact of adversarial perturbations on context-dependent data, we design a contextual adversarial training (CAT) strategy to learn more diverse features from context and enhance the model's context robustness. Under the framework with CAT, we develop a sequence-based SACL-LSTM to learn label-consistent and context-robust features for ERC. Experiments on three datasets show that SACL-LSTM achieves state-of-the-art performance on ERC. Extended experiments prove the effectiveness of SACL and CAT.

1 Introduction

Emotion recognition in conversations (ERC) aims to detect emotions expressed by speakers during a conversation. The task is a crucial topic for developing empathetic machines (Ma et al., 2020). Existing works mainly focus on context modeling (Majumder et al., 2019; Ghosal et al., 2019; Hu et al., 2021a) and emotion representation learning (Zhu et al., 2021; Yang et al., 2022; Li et al., 2022a) to recognize emotions. However, these methods have limitations in discovering the intrinsic structure of data relevant to emotion labels, and struggle to extract generalized and robust representations, resulting in mediocre recognition performance.

In the field of representation learning, label-based contrastive learning (Khosla et al., 2020; Martín et al., 2022) techniques are used to learn a generalized representation by capturing similarities between examples within a class and contrasting them with examples from other classes. Since similar emotions often have similar context and overlapping feature spaces, these techniques that directly compress the feature space of each class are likely to hurt the fine-grained features of each emotion, thus limiting the ability of generalization.

To address these, we propose a supervised adversarial contrastive learning (SACL) framework to learn class-spread structured representations in a supervised manner. SACL applies contrast-aware adversarial training to generate worst-case samples and uses a joint class-spread contrastive learning objective on both original and adversarial samples. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features.

Specifically, we adopt soft[1] SCL (Gunel et al., 2021) on original samples to obtain contrast-aware adversarial perturbations. Then, we put perturbations on the hidden layers to generate hard positive examples with a min-max training recipe. These generated samples can spread out the representation space for each class and confuse robust-less networks. After that, we utilize a new soft SCL on obtained adversarial samples to maximize the consistency of class-spread representations with the same label. Under the joint objective on both original and adversarial samples, the network can effectively learn label-consistent features and achieve better generalization.

In context-dependent dialogue scenarios, directly generating adversarial samples interferes with the correlation between utterances, which is detrimental to context understanding. To avoid this, we design a contextual adversarial training (CAT) strategy to adaptively generate context-level

---

[*]Corresponding author.

[1]The soft version means a cross-entropy term is added to alleviate the class collapse issue (Graf et al., 2021), wherein each point in the same class has the same representation.

worst-case samples and extract more diverse features from context. This strategy applies adversarial perturbations to the context-aware network structure in a multi-channel way, instead of directly putting perturbations on context-free layers in a single-channel way (Goodfellow et al., 2015; Miyato et al., 2017). After introducing CAT, SACL can further learn more diverse features and smooth representation spaces from context-dependent inputs, as well as enhance the model's context robustness.

Under SACL framework, we design a sequence-based method SACL-LSTM to recognize emotion in the conversation. It consists of a dual long short-term memory (Dual-LSTM) module and an emotion classifier. Dual-LSTM is a modified version of the contextual perception module (Hu et al., 2021a), which can effectively capture contextual features from a dialogue. With the guidance of SACL, the model can learn label-consistent and context-robust emotional features for the ERC task.

We conduct experiments on three public benchmark datasets. Results consistently demonstrate that our SACL-LSTM significantly outperforms other state-of-the-art methods on the ERC task, showing the effectiveness and superiority of our method. Moreover, extensive experiments prove that our SACL framework can capture better structured and robust representations for classification.

The main contributions are as follows: **1)** We propose a supervised adversarial contrastive learning (SACL) framework to extract class-spread structured representations for classification. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features. **2)** We design a contextual adversarial training (CAT) strategy to learn more diverse features from context-dependent inputs and enhancing the model's context robustness. **3)** We develop a sequence-based method SACL-LSTM under the framework to learn label-consistent and context-robust emotional features for ERC[2]. **4)** Experiments on three benchmark datasets show that SACL-LSTM significantly outperforms other state-of-the-art methods, and prove the effectiveness of the SACL framework[3].

2 Methodology

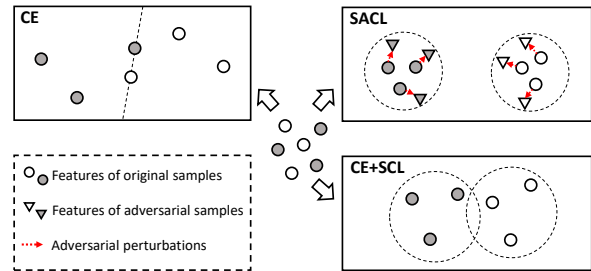In this section, we first present the methodology of SACL framework. Besides, for better adaptation to

Figure 1: Comparison of different training objectives on a two-class case. CE and CE+SCL mean cross-entropy and soft supervised contrastive learning, respectively.

context-independent scenarios, we introduce a CAT strategy to SACL framework. Finally, we apply the proposed SACL framework for emotion recognition in conversations and provide a sequence-based method SACL-LSTM.

2.1 Supervised Adversarial Contrastive Learning Framework

In the field of representation learning, label-based contrastive learning (Khosla et al., 2020; Martín et al., 2022) techniques are used to learn a generalized representation by capturing similarities between examples within a class and contrasting them with examples from other classes. However, directly compressing the feature space of each class is prone to harming fine-grained intra-class features, which limits the model's ability to generalize.

To address this, we design a supervised adversarial contrastive learning (SACL) framework for learning class-spread structured representations. The framework applies contrast-aware adversarial training to generate worst-case samples and uses a joint class-spread contrastive learning objective on both original and adversarial samples. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features. Figure 1 visualizes the difference between SACL and two representative optimization objectives (i.e., CE and soft SCL (Gunel et al., 2021)) on a toy example.

Formally, let us denote I as the set of samples in a mini-batch. Define $\phi(i) = \{e \in I \backslash \{i\} : \hat{\mathbf{y}}_e = \hat{\mathbf{y}}_i\}$ is the set of indices of all positives in the mini-batch distinct from i, and $|\phi(i)|$ is its cardinality. The loss function of soft SCL is a weighted average of CE loss and SCL loss with a trade-off scalar parameter λ, i.e.,

$$\mathcal{L}_{\text{soft-SCL}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{SCL}}, \qquad (1)$$

---

[2]To the best of our knowledge, this is the first attempt to introduce the idea of adversarial training into the ERC task.

[3]The source code is available at https://github.com/zerohd4869/SACL

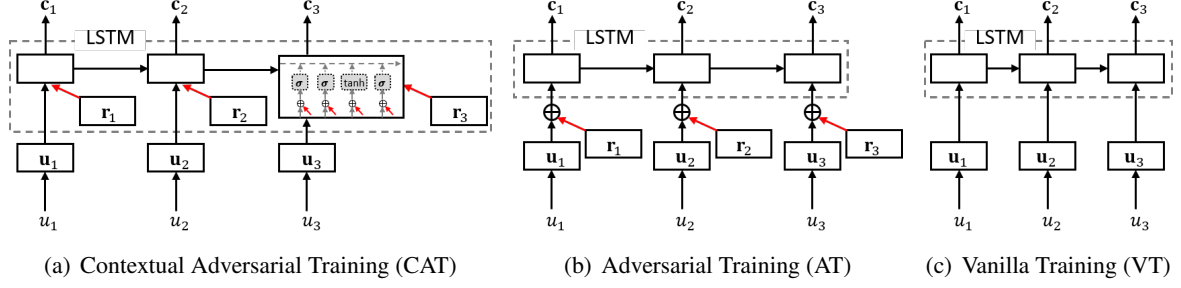(a) Contextual Adversarial Training (CAT)   (b) Adversarial Training (AT)   (c) Vanilla Training (VT)

Figure 2: An LSTM network against different perturbations. CAT, AT and VT represent the hidden layers with contextual adversarial perturbations $r_{\text{c-adv}}$, adversarial perturbations r_{adv} and no perturbations, respectively.

where

$$\mathcal{L}_{\text{CE}} = -\sum_{i \in I} \mathbf{y}_{i,k} \log(\hat{\mathbf{y}}_{i,k}), \tag{2}$$

$$\mathcal{L}_{\text{SCL}} = \sum_{i \in I} \frac{-1}{|\phi(i)|} \sum_{e \in \phi(i)} \log \frac{\exp(sim(\mathbf{z}_i, \mathbf{z}_e)/\tau)}{\sum_{a \in A(i)} \exp(sim(\mathbf{z}_i, \mathbf{z}_a)/\tau)}, \tag{3}$$

$\mathbf{y}_{i,k}$ and $\hat{\mathbf{y}}_{i,k}$ denote the value of one-hot vector $\mathbf{y}_i$ and probability vector $\hat{\mathbf{y}}_i$ at class index k, respectively. $A(i) = I \backslash \{i\}$. $\mathbf{z}_i$ refers to the hidden representation of the network's output for the i-th sample. $sim(\cdot, \cdot)$ is a pairwise similarity function, i.e., dot product. $\tau > 0$ is a scalar temperature parameter that controls the separation of classes.

At each step of training, we apply an adversarial training strategy with the soft SCL objective on original samples to produce anti-contrast worst-case samples. The training strategy can be implemented using a context-free approach such as FGM (Miyato et al., 2017) or our context-aware CAT. These samples can be seen as hard positive examples, which spread out the representation space for each class and confuse the robust-less model. After that, we utilize a new soft SCL on obtained adversarial samples to maximize the consistency of class-spread representations with the same label. Following the above calculation process of $\mathcal{L}_{\text{soft-SCL}}$ on original samples, the optimization objective on corresponding adversarial samples can be easily obtained in a similar way, i.e., $\mathcal{L}_{\text{soft-SCL}}^{\text{r-adv}}$.

The overall loss of SACL is defined as a sum of two soft SCL losses on both original and adversarial samples, i.e.,

$$\mathcal{L} = \mathcal{L}_{\text{soft-SCL}} + \mathcal{L}_{\text{soft-SCL}}^{\text{r-adv}}. \tag{4}$$

2.2 Contextual Adversarial Training

Adversarial training (AT) (Goodfellow et al., 2015; Miyato et al., 2017) is a widely used regularization method for models to improve robustness to small, approximately worst-case perturbations. In

context-dependent scenarios, directly generating adversarial samples interferes with the correlation between samples, which is detrimental to context understanding.

To avoid this, we design a contextual adversarial training (CAT) strategy for a context-aware network, to obtain diverse context features and a robust model. Different from the standard AT that put perturbations on context-free layers (e.g., word/sentence embeddings), we add adversarial perturbations to the context-aware network structure in a multi-channel way. Under a supervised training objective, it can obtain diverse features from context and enhance model robustness to contextual perturbations.

Let us denote (u, y) as the mini-batch input sampled from distribution D and $p(y|u; \theta)$ as a context-aware model. At each step of training, we identify the contextual adversarial perturbations $r_{\text{c-adv}}$ against the current model with the parameters $\hat{\theta}$, and put them on the context-aware hidden layers of the model. With a linear approximation (Goodfellow et al., 2015), an L_q norm-ball and a certain radius ϵ for $r_{\text{c-adv}}$, and a training objective ℓ (e.g., soft SCL), the formulation of CAT is illustrated by

$$\min_{\theta} \mathbb{E}_{(u,y) \sim D} \max_{\|r_{\text{c-adv}}\|_q \leq \epsilon} \ell(u + r_{\text{c-adv}}, y; \theta),$$
$$\text{where } r_{\text{c-adv}} = -\epsilon g/\|g\|_q, \ g = \nabla_u \log p(y \mid u; \hat{\theta}). \tag{5}$$

Here, we take the LSTM network (Hochreiter and Schmidhuber, 1997) with a sequence input $[u_1, u_2, ..., u_N]$ as an example, and the corresponding representations of the output are $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N]$. Adversarial perturbations are put on context-aware hidden layers of the LSTM in a multi-channel way, including three gated layers and a memory cell layer in the LSTM structure, as shown in Figure 2.

With contextual perturbations on the network, there is a reasonable interpretation of the formulation in Eq. (5). The inner maximization problem
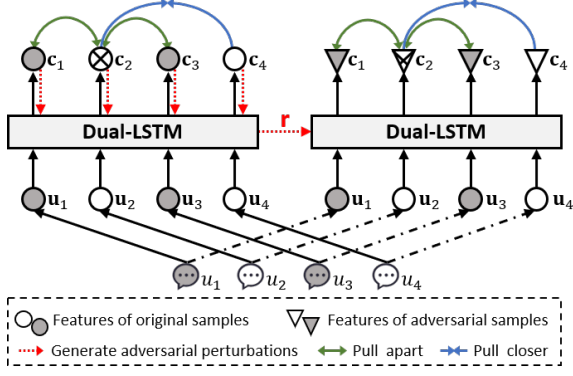
Figure 3: The overall architecture of SACL-LSTM. The mark with/without shade means two different classes. We take the $\times$-marked utterance as an example to show the objective of SACL. Dual-LSTM is a context-aware module that can capture sequential features in the conversation. r means contextual adversarial perturbations that put on hidden layers of Dual-LSTM.

is finding the context-level worst-case samples for the network, and the outer minimization problem is to train a robust network to the worst-case samples. After introducing CAT, our SACL can further learn more diverse features and smooth representation spaces from context-dependent inputs, as well as enhance the model's context robustness.

2.3 Application for Emotion Recognition in Conversations

In this subsection, we apply SACL framework to the task of emotion recognition in conversations (ERC), and present a sequence-based method SACL-LSTM. The overall architecture is illustrated in Figure 3. With the guidance of SACL with CAT, the method can learn label-consistent and context-robust emotional features for better emotion recognition.

2.3.1 Problem Statement

The ERC task aims to recognize emotions expressed by speakers in a conversation. Formally, let $U = [u_1, u_2, ..., u_N]$ be a conversation with N utterances and M speakers/parties. Each utterance u_i is spoken by the party $p_{\phi(u_i)} \in \{p_1, p_2, ..., p_M\}$, where ϕ maps the utterance index into the corresponding speaker index. For each $m \in [1, M]$, U_m represents the set of utterances spoken by the party p_m, i.e., $U_m = \{u_i | u_i \in U$ and u_i is spoken by $p_m, \forall i \in [1, N]\}$. The goal is to identify the emotion label y_i for each utterance u_i from the pre-defined emotions $\mathcal{Y}$.

2.3.2 Textual Feature Extraction

Following previous works (Ghosal et al., 2020; Shen et al., 2021b), the pre-trained *roberta-large*[4] (Liu et al., 2019) is fine-tuned on the train sets for utterance-level emotion classification, and then its parameters are frozen when training our model. Formally, given an utterance input u_i, the output of $[CLS]$ token in the last hidden layer of the encoder is used to obtain the utterance representation $\mathbf{u}_i$ with a dimension d_u. We denote $\{\mathbf{u}_i \in \mathbb{R}^{d_u}\}_{i=1}^N$ as the context-free textual features for N utterances.

2.3.3 Model Structure

The network structure of SACL-LSTM consists of a dual long short-term memory (Dual-LSTM) module and an emotion classifier.

Dual-LSTM After extracting textual features, we design a Dual-LSTM module to capture situation- and speaker-aware contextual features in a conversation. It is a modified version of the contextual perception module in Hu et al. (2021a).

Specifically, to alleviate the speaker cold-start issue[5], we modify the speaker perception module. If the number of utterances of the speaker is less than a predefined integer threshold ξ, the common characteristics of these cold-start speakers are directly represented by a shared general speaker vector $\mathbf{o}$. The speaker-aware features $\mathbf{c}_i^{sp}$ are computed as:

$$\mathbf{c}_i^{sp}, \mathbf{h}_{m,j}^{sp} = \begin{cases} \mathbf{o}, \text{None} & \text{if } |U_m| < \xi, \\ \overrightarrow{LSTM}^{sp}(\mathbf{u}_i, \mathbf{h}_{m,j-1}^{sp}), j \in [1, |U_m|] & \text{otherwise}, \end{cases} \tag{6}$$

where $\overleftrightarrow{LSTM}^{sp}$ indicates a BiLSTM to obtain speaker embeddings. $\mathbf{h}_{m,j}^{sp}$ is the j-th hidden state of the party p_m with a dimension of d_h. $m = \phi(u_i)$. U_m refers to all utterances of p_m in a conversation. The situation-aware features $\mathbf{c}_i^{si}$ are defined as,

$$\mathbf{c}_i^{si}, \mathbf{h}_i^{si} = \overleftrightarrow{LSTM}^{si}(\mathbf{u}_i, \mathbf{h}_{i-1}^{si}), \tag{7}$$

where $\overleftrightarrow{LSTM}^{si}$ is a BiLSTM to obtain situation-aware embeddings and $\mathbf{h}_i^{si}$ is the hidden vector with a dimension of d_h.

We concatenate the situation-aware and speaker-aware features to form the context representation of each utterance, i.e., $\mathbf{c}_i = [\mathbf{c}_i^{si}; \mathbf{c}_i^{sp}]$.

Emotion Classifier Finally, according to the context representation, an emotion classifier is applied to predict the emotion label of each utterance.

$$\hat{\mathbf{y}}_i = softmax(\mathbf{W}_c \mathbf{c}_i + \mathbf{b}_c), \qquad (8)$$

where $\mathbf{W}_c \in \mathbb{R}^{4d_h \times |\mathcal{Y}|}$ and $\mathbf{b}_c \in \mathbb{R}^{|\mathcal{Y}|}$ are trainable parameters. $|\mathcal{Y}|$ is the number of emotion labels.

2.3.4 Optimization Process

Under SACL framework, we apply contrast-aware CAT to generate worst-case samples and utilize a joint class-spread contrastive learning objective on both original and adversarial samples. At each step of training, we apply the CAT strategy with the soft SCL objective on original samples to produce context-level adversarial perturbations. The perturbations are put on context-aware hidden layers of Dual-LSTM in a multi-channel way, and then obtain adversarial samples. After that, we leverage a new soft SCL on these worst-case samples to maximize the consistency of emotion-spread representations with the same label. Under the joint objective on both original and adversarial samples, SACL-LSTM can learn label-consistent and context-robust emotional features for ERC.

3 Experimental Setups

3.1 Datasets

We evaluate our model on three benchmark datasets. **IEMOCAP** (Busso et al., 2008) contains dyadic conversation videos between pairs of ten unique speakers, where the first eight speakers belong to train sets and the last two belong to test sets. The utterances are annotated with one of six emotions, namely happy, sad, neutral, angry, excited, and frustrated. **MELD** (Poria et al., 2019a) contains multi-party conversation videos collected from Friends TV series. Each utterance is annotated with one of seven emotions, i.e., joy, anger, fear, disgust, sadness, surprise, and neutral. **EmoryNLP** (Zahiri and Choi, 2018) is a textual corpus that comprises multi-party dialogue transcripts of the Friends TV show. Each utterance is annotated with one of seven emotions, i.e., sad, mad, scared, powerful, peaceful, joyful, and neutral.

The statistics are reported in Table 1. In this paper, we focus on ERC in a textual setting. Other multimodal knowledge (i.e., acoustic and visual modalities) is not used. We use the pre-defined train/val/test splits in MELD and EmoryNLP. Following previous studies (Hazarika et al., 2018b;

Dataset	# Dialogues			# Utterances			# Avg. Turns	# Avg. Parties
	train	val	test	train	val	test		
IEMOCAP	120		31	5810		1623	49.2	2
MELD	1039	114	280	9989	1109	2610	9.6	2.7
EmoryNLP	659	89	79	7551	954	984	11.5	3.2

Table 1: The statistics of three datasets.

Ghosal et al., 2019), we randomly extract 10% of the training dialogues in IEMOCAP as validation sets since there is no predefined train/val split.

3.2 Comparison Methods

The fourteen baselines compared are as follows. 1) Sequence-based methods: **bc-LSTM** (Poria et al., 2017) employs an utterance-level LSTM to capture contextual features. **DialogueRNN** (Majumder et al., 2019) is a recurrent network to track speaker states and context. **COSMIC** (Ghosal et al., 2020) uses GRUs to incorporate commonsense knowledge and capture complex interactions. **Dialogue-CRN** (Hu et al., 2021a) is a cognitive-inspired network with multi-turn reasoning modules that captures implicit emotional clues in a dialogue. **CauAIN** (Zhao et al., 2022) uses causal clues in commonsense knowledge to enrich the modeling of speaker dependencies.

2) Graph-based methods: **DialogueGCN** (Ghosal et al., 2019) uses GRUs and GCNs with relational edges to capture context and speaker dependency. **RGAT** (Ishiwatari et al., 2020) applies position encodings to RGAT to consider speaker and sequential dependency. **DAG-ERC** (Shen et al., 2021b) adopts a directed GNN to model the conversation structure. **SGED+DAG** (Bao et al., 2022) is a speaker-guided framework with a one-layer DAG that can explore complex speaker interactions.

3) Transformer-based methods: **KET** (Zhong et al., 2019) incorporates commonsense knowledge and context into a Transformer. **DialogXL** (Shen et al., 2021a) adopts a modified XLNet to deal with longer context and multi-party structures. **TOD-KAT** (Zhu et al., 2021) enhances the ability of Transformer by incorporating commonsense knowledge and a topic detection task. **CoG-BART** (Li et al., 2022a) uses a SupCon loss (Khosla et al., 2020) and a response generation task to enhance BART's ability. **SPCL+CL** (Song et al., 2022) applies a prompt-based BERT with supervised prototypical contrastive learning (Wang et al., 2021; Martín et al., 2022) and curriculum learning (Bengio et al., 2009).

Methods	# Param.	IEMOCAP		MELD		EmoryNLP		Avg.	
		Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
Transformer-based Methods									
KET[†‡] (Zhong et al., 2019)	-	-	59.56	-	58.18	-	34.39	-	50.17
DialogXL[‡] (Shen et al., 2021a)	-	-	65.94	-	62.41	-	34.73	-	54.36
TODKAT[†‡] (Zhu et al., 2021)	-	61.11	61.33	67.24	65.47	42.38	38.69	56.91	55.16
CoG-BART (Li et al., 2022a)	415.1M	65.02	64.87	64.95	63.82	40.94	37.33	56.97	55.34
SPCL+CL (Song et al., 2022)	356.7M	66.71	66.93	64.36	64.93	40.32	39.45	57.13	57.10
Graph-based Methods									
DialogueGCN (Ghosal et al., 2019)	2.1M	62.49	62.11	63.62	62.68	36.87	34.63	54.33	53.14
RGAT[‡] (Ishiwatari et al., 2020)	-	-	65.22	-	60.91	-	34.42	-	53.52
DAG-ERC (Shen et al., 2021b)	9.5M	66.54	66.53	63.75	63.36	39.64	38.29	56.64	56.06
SGED+DAG (Bao et al., 2022)	3.0M	66.29	66.27	63.60	63.16	39.19	38.11	56.36	55.85
Sequence-based Methods									
bc-LSTM (Poria et al., 2017)	1.2M	63.08	62.84	65.87	64.87	40.85	36.84	56.60	54.85
DialogueRNN (Majumder et al., 2019)	9.9M	64.85	64.65	65.96	65.30	**43.66**	37.54	58.16	55.83
COSMIC[†] (Ghosal et al., 2020)	11.9M	63.43	63.43	65.96	65.03	41.79	38.49	57.06	55.65
DialogueCRN (Hu et al., 2021a)	3.3M	67.39	67.53	66.93	65.77	41.04	38.79	58.45	57.36
CauAIN[†] (Zhao et al., 2022)	6.1M	65.08	65.01	65.85	64.89	43.13	37.87	58.02	55.92
SACL-LSTM (ours)	2.6M	**69.08***	**69.22***	**67.51***	**66.45***	42.21	**39.65***	**59.60***	**58.44***

Table 2: Overall results (%) against various methods for ERC. We present accuracy (Acc) and weighted-F1 (w-F1) score for each dataset. † means the external knowledge is used. # Param. means the average number of learnable model parameters. ‡ means the results are from the original paper or their official repository; results of CoG-BART and SPCL+CL are reproduced under model initialization with *bart-large*[4] and *roberta-large*[4], respectively; all other results are reproduced using *roberta-large* features that our SACL-LSTM uses. For each reproduced method, we run five random seeds and report the average result on test sets. Best results are highlighted in bold. * represents statistical significance over state-of-the-art scores under the paired t-test (p<0.05).

3.3 Evaluation Metrics

Following previous works (Hu et al., 2021a; Li et al., 2022a), we report the accuracy and weighted-F1 score to measure the overall performance. Also, the F1 score per class and macro-F1 score are reported to evaluate the fine-grained performance. For the structured representation evaluation, we choose three supervised clustering metrics (i.e., ARI, NMI, and FMI) and three unsupervised clustering metrics (i.e., SC, CHI, and DBI) to measure the clustering performance of learned representations. For the empirical robust evaluation (Carlini and Wagner, 2017), we use the robust weighted-F1 score on adversarial samples generated from original test sets. Besides, the paired t-test (Kim, 2015) is used to verify the statistical significance of the differences between the two approaches.

3.4 Implementation Details

All experiments are conducted on a single NVIDIA Tesla V100 32GB card. The validation sets are used to tune hyperparameters and choose the optimal model. For each method, we run five random seeds and report the average result of the test sets. The network parameters of our model are optimized by using Adam optimizer (Kingma and Ba, 2015). More experimental details are listed in Appendix B.

(a) IEMOCAP

Methods	*Hap.*	*Sad.*	*Neu.*	*Ang.*	*Exc.*	*Fru.*	*Avg.*
DialogueCRN	54.28	81.34	69.57	62.09	67.33	64.22	66.47
SACL-LSTM	**56.91***	**84.78***	**70.00***	**64.09***	**69.70***	**65.02***	**68.42***
Improve	+2.63	+3.44	+0.43	+2.00	+2.37	+0.80	+1.95

(b) MELD

Methods	*Neu.*	*Sur.*	*Fea.*	*Sad.*	*Joy.*	*Dis.*	*Ang.*	*Avg.*
DialogueCRN	79.72	57.62	18.26	39.30	64.56	**32.07**	**52.53**	49.15
SACL-LSTM	**80.17***	**58.77***	**26.23***	**41.34***	**64.98***	31.47	52.35	**50.76***
Improve	+0.45	+1.15	+7.97	+2.04	+0.42	-0.60	-0.18	+1.61

(c) EmoryNLP

Methods	*Joy.*	*Mad.*	*Pea.*	*Neu.*	*Sad.*	*Pow.*	*Sca.*	*Avg.*
DialogueCRN	54.42	36.44	10.18	53.83	25.74	4.55	**37.49**	31.81
SACL-LSTM	**54.78***	**37.68***	**11.66***	**55.42***	**25.83**	**5.43***	37.11	**32.56***
Improve	+0.36	+1.24	+1.48	+1.59	+0.09	+0.88	-0.38	+0.75

Table 3: Fine-grained results (%) of SACL-LSTM and DialogueCRN for all emotion categories. DialogueCRN is the sub-optimal method in Table 2. We report F1 score per class and macro-F1 score.

4 Results and Analysis

4.1 Overall Results

The overall results[6] are reported in Table 2. SACL-LSTM consistently obtains the best weighted-F1 score over comparison methods on three datasets. Specifically, SACL-LSTM obtains **+1.1%** absolute

---

[6]We noticed that DialogueRNN and CauAIN present a poor weighted-F1 but a fine accuracy score on EmoryNLP, which is most likely due to the highly class imbalance issue.

Methods	IEMOCAP	MELD	EmoryNLP
SACL-LSTM	**69.22**$_{\pm0.54}$	**66.45**$_{\pm0.35}$	**39.65**$_{\pm0.66}$
- w/o SACL	68.17$_{\pm0.63}$	65.64$_{\pm0.14}$	38.65$_{\pm0.62}$
- w/o SACL - w/o Dual-LSTM	52.99$_{\pm0.49}$	64.65$_{\pm0.12}$	37.74$_{\pm0.20}$

Table 4: Ablation results (%) of SACL-LSTM. "- w/o SACL" means replacing the SACL with a cross-entropy term. "- w/o Dual-LSTM" means replacing the Dual-LSTM with an MLP. We report the average score and standard deviation of the weighted-F1 with five seeds.

improvements over other state-of-the-art methods in terms of the average weighted-F1 score on three datasets. Besides, SACL-LSTM obtains **+1.2%** absolute improvements in terms of the average accuracy score. The results indicates the good generalization ability of our method to unseen test sets.

We also report fine-grained results on three datasets in Table 3. SACL-LSTM achieves better results for most emotion categories (17 out of 20 classes), except three classes (i.e., disgust and anger in MELD, and scared in EmoryNLP). It is worth noting that SACL-LSTM obtains **+2.0%**, **+1.6%** and **+0.8%** absolute improvements in terms of the macro-F1 (average score of F1 for all classes) on IEMOCAP, MELD and EmoryNLP, respectively.

4.2 Ablation Study

We conduct ablation studies to evaluate key components in SACL-LSTM. The results are shown in Table 4. When removing the proposed SACL framework (i.e., - w/o SACL) and replacing it with a simple cross-entropy objective, we obtain inferior performance in terms of all metrics. When further removing the context-aware Dual-LSTM module (i.e., - w/o SACL - w/o Dual-LSTM) and replacing it with a context-free MLP (i.e., a fully-connected neural network with a single hidden layer), the results decline significantly on three datasets. It shows the effectiveness of both components.

4.3 Comparison with Different Optimization Objectives

To demonstrate the superiority of SACL, we include control experiments that replace it with the following optimization objectives, i.e., CE+SCL (soft SCL) (Gunel et al., 2021), CE+SupCon[7] (Khosla et al., 2020), and cross-entropy (CE).

Table 5 shows results against various optimization objectives. SACL significantly outperforms the comparison objectives on three datasets.

---

[7]The idea of SupCon is very similar to SCL. Their implementations are slightly different. Combined with CE, they achieved very close performance, as shown in Table 5.

Optimization Objectives	IEMOCAP	MELD	EmoryNLP
SACL	**69.22**$_{\pm0.54}$	**66.45**$_{\pm0.35}$	**39.65**$_{\pm0.66}$
CE+SCL	68.32$_{\pm0.45}$	65.95$_{\pm0.20}$	38.93$_{\pm0.89}$
CE+SupCon	68.37$_{\pm0.36}$	65.89$_{\pm0.38}$	39.00$_{\pm0.93}$
CE	68.17$_{\pm0.63}$	65.64$_{\pm0.14}$	38.65$_{\pm0.62}$

Table 5: Comparison results (%) against different optimization objectives. We report the weighted-F1 score.

Training Strategies	IEMOCAP	MELD	EmoryNLP
SACL			
- w/ **CAT**	**69.22**$_{\pm0.54}$	**66.45**$_{\pm0.35}$	**39.65**$_{\pm0.66}$
- w/ CRT	68.28$_{\pm0.72}$	65.70$_{\pm0.29}$	39.16$_{\pm0.59}$
- w/ AT	68.95$_{\pm1.03}$	65.69$_{\pm0.34}$	38.58$_{\pm0.39}$
- w/ VT	68.32$_{\pm0.45}$	65.95$_{\pm0.20}$	38.93$_{\pm0.89}$

Table 6: Comparison results (%) against different training strategies under the SACL framework. CAT, CRT, AT, and VT are contextual adversarial training, contextual random training, adversarial training, and vanilla training, respectively. We report the weighted-F1 score.

CE+SCL and CE+SupCon objectives apply label-based contrastive learning to extract a generalized representation, leading to better performance than CE. However, they compress the feature space of each class and harm fine-grained intra-class features, yielding inferior results than our SACL. SACL uses a joint class-spread contrastive learning objective on both original and adversarial samples. It can effectively utilize label-level feature consistency and retain fine-grained intra-class features.

4.4 Comparison with Different Training Strategies

To evaluate the effectiveness of contextual adversarial training (CAT), we compare with different training strategies, i.e., adversarial training (AT) (Miyato et al., 2017), contextual random training (CRT), and vanilla training (VT). CRT is the strategy in which we replace $r_{\text{c-adv}}$ in CAT with random perturbations from a multivariate Gaussian with the scaled norm on context-aware hidden layers.

The results are reported in Table 6. Compared with other strategies, our CAT obtains better performance consistently on three datasets. It shows that CAT can enhance the diversity of emotional features by adding adversarial perturbations to the context-aware structure with a min-max training recipe. We notice that AT strategy achieves the worst performance on MELD and EmoryNLP with the extremely short length of conversations. It indicates that AT is difficult to improve the diversity of context-dependent features with a limited context.

	Optimization Objectives	Supervised Metrics			Unsupervised Metrics		
		ARI↑ (%)	NMI↑ (%)	FMI↑ (%)	SC↑	CHI↑	DBI↓
IEMOCAP	CE	40.61	47.39	51.56	0.36	1231.19	1.06
	CE+SCL	40.55	47.25	51.53	0.36	1188.46	1.07
	SACL	**41.95**	**48.26**	**52.62**	**0.39**	**1696.05**	**0.99**
	Improve	+1.34	+0.87	+1.06	+0.03	+464.86	+0.07
MELD	CE	40.74	27.00	59.41	0.24	755.83	1.41
	CE+SCL	40.92	27.13	59.52	0.23	738.70	1.42
	SACL	**42.34**	**28.22**	**60.42**	**0.31**	**1342.38**	**1.16**
	Improve	+1.42	+1.09	+0.90	+0.07	+586.55	+0.25

Table 7: Clustering results against different optimization objectives. Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Index (FMI) evaluate the accuracy of clustering. Silhouette Coefficient (SC), Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI) evaluate the separation and compactness of clustering. SC, CHI, and DBI are evaluated based on K-Means, and we define the number of clusters K as the true number of categories, i.e., $K = 6$ for IEMOCAP, and $K = 7$ for MELD.

4.5 Structured Representation Evaluation

To evaluate the quality of structured representations, we measure the clustering performance based on the representations learned with different optimization objectives on the test set of IEMOCAP and MELD. Table 7 reports the clustering results of the Dual-LSTM network under three optimization objectives, including CE, CE+SCL, and our SACL.

According to supervised clustering metrics, the proposed SACL outperforms other optimization objectives by **+1.3%** and **+1.4%** in ARI, **+0.9%** and **+1.1%** in NMI, **+1.1%** and **+0.9%** in FMI for IEMOCAP and MELD, respectively. The more accurate clustering results show that our SACL can distinguish different data categories and assign similar data points to the same categories. It indicates that SACL can discover the intrinsic structure of data relevant to labels and extract generalized representations for emotion recognition.

According to unsupervised clustering metrics, SACL achieves better results than other optimization objectives by **+0.03** and **+0.07** in SC, **+464.86** and **+586.55** in CHI, and **+0.07** and **+0.25** in DBI for IEMOCAP and MELD, respectively. Better performance on these metrics suggests that SACL can learn more clear, separated, and compact clusters. This indicates that SACL can better capture the underlying structure of the data, which can be beneficial for subsequent emotion recognition.

Overall, the results demonstrate the effectiveness of the SACL framework in learning structured representations for improving clustering performance and quality, as evidenced by the significant im-
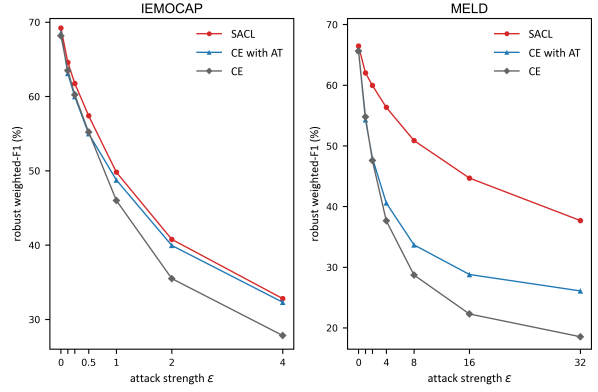

Figure 4: Context robustness performances against different optimization objectives. We report the robust weighted-F1 scores under different attack strengths. Detailed results are listed in Appendix C.1.

provements in various clustering metrics.

4.6 Context Robustness Evaluation

We further validate context robustness against different optimization objectives. We adjust different attack strengths of CE-based contextual adversarial perturbations on the test set and report the robust weighted-F1 scores. The context robustness results of SACL, CE with AT, and CE objectives on IEMOCAP and MELD are shown in Figure 4. CE with AT means using a cross-entropy objective with traditional adversarial training, i.e., FGM.

Our SACL consistently gains better robust weighted-F1 scores over other optimization objectives on both datasets. Under different attack strengths ($\epsilon > 0$), SACL-LSTM achieves up to **2.2%** (average **1.3%**) and **17.2%** (average **13.4%**) absolute improvements on IEMOCAP and MELD, respectively. CE with AT obtains sub-optimal performance since generating context-free adversarial samples interferes with the correlation between utterances, which is detrimental to context understanding. Our SACL using CAT can generate context-level worst-case samples for better training and enhance the model's context robustness.

Moreover, we observe that SACL achieves a significant improvement on MELD with limited context. The average number of dialogue turns in MELD is relatively small, making it more likely for any two utterances to be strongly correlated. By introducing CAT, SACL learns more diverse features from the limited context, obtaining better context robustness results on MELD than others.
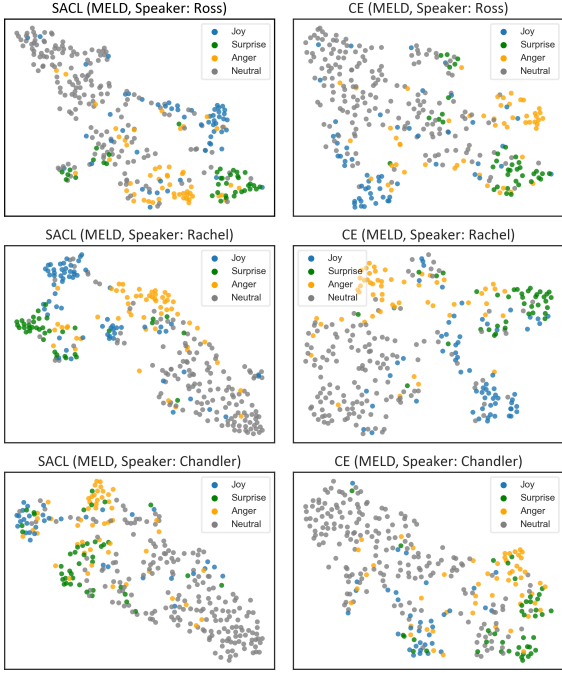
Figure 5: t-SNE visualization of representations learned with different optimization objectives on MELD. The data points reflect an overall distribution representation. Points corresponding to categories with a sample proportion of less than 10% are excluded for a clear picture.

4.7 Representation Visualization

We qualitatively visualize the learned representations on the test set of MELD with t-SNE (Van der Maaten and Hinton, 2008). Figure 5 shows the visualization of the three speakers. Compared with using CE objective, the distribution of each emotion class learned by our SACL is more tight and united. It indicates that SACL can learn cluster-level structured representations and have a better ability to generalization. Besides, under SACL, the representations of surprise are away from neutral, and close to both joy and anger, which is consistent with the nature of surprise[8]. It reveals that SACL can partly learn inter-class intrinsic structure in addition to intra-class feature consistency.

4.8 Error Analysis

Figure 6 shows an error analysis of SACL-LSTM and its ablated variant on the test set of IEMOCAP and MELD. The normalized confusion matrices are used to evaluate the quality of each model's predicted outputs. From the diagonal elements of the matrices, SACL-LSTM reports better true positives against others on most fine-grained emotion cate-

---

[8]Surprise is a non-neutral complex emotion that can be expressed with positive or negative valence (Poria et al., 2019a).
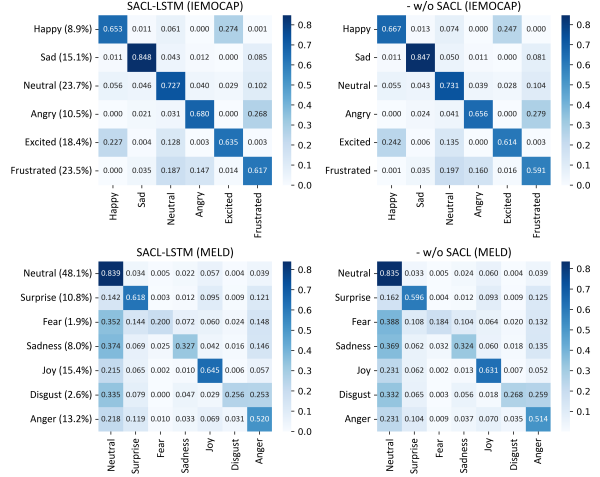

Figure 6: The normalized confusion matrices for SACL-LSTM and its variant. The rows represent the actual classes, whereas the columns represent predictions made by the model. Each cell (i, j) represents the percentage of class i predicted to be class j.

gories. It suggests that SACL-LSTM is unbiased towards the under-represented emotion labels and learns better fine-grained features. Compared with the ablated variant w/o SACL, SACL-LSTM obtains better performances at similar categories, e.g., excited to happy, angry to frustrated, and frustrated to angry on IEMOCAP. It indicates that the SACL framework can effectively mitigate the misclassification problem of similar emotions. The poor effect of happy to excited may be due to the small proportion of happy samples used for training. For MELD, some categories (i.e., fear, sadness, and disgust) that account for a small proportion are easily misclassified as neutral accounting for nearly half, which is caused by the class imbalance issue.

5 Conclusion

We propose a supervised adversarial contrastive learning framework to learn class-spread structured representations for classification. It applies a contrast-aware adversarial training strategy and a joint class-spread contrastive learning objective. Besides, we design a contextual adversarial training strategy to learn more diverse features from context-dependent inputs and enhance the model's context robustness. Under the SACL framework with CAT, we develop a sequence-based method SACL-LSTM to learn label-consistent and context-robust features on context-dependent data for better emotion recognition. Experiments verified the effectiveness of SACL-LSTM for ERC and SACL for learning generalized and robust representations.

Limitations

In this paper, we present a supervised adversarial contrastive learning (SACL) framework with contextual adversarial training to learn class-spread structured representations for context-dependent emotion classification. However, the framework is somewhat limited by the class imbalance issue, as illustrated in Section 4. To more comprehensively evaluate the generalization of SACL, it is necessary to test its transferability in low-resource and out-of-distribution scenarios, and evaluate its performance across a wider range of tasks. Additionally, it would be beneficial to explore the theoretical underpinnings and potential applications of the framework in greater depth. The aforementioned limitations will be left for future research.

Acknowledgements

References

Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15509–15519.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4051–4057. ijcai.org.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

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.

Nicholas Carlini and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. 2021. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21480–21492.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2470–2481. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Florian Graf, Christoph D. Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised constrastive learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3821–3830. PMLR.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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, 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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132. Association for Computational Linguistics.

Olivier J. Hénaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Dou Hu, Xiaolong Hou, Xiyang Du, Mengyuan Zhou, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022a. VarMAE: Pre-training of variational masked autoencoder for domain-adaptive language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6276–6286. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022b. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041. IEEE.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7042–7052. Association for Computational Linguistics.

Dou Hu, Mengyuan Zhou, Xiyang Du, Mengfei Yuan, Jin Zhi, Lian-Xin Jiang, Yang Mo, and Xiaofeng Shi. 2022c. PALI-NLP at SemEval-2022 Task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 335–343. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5666–5675. Association for Computational Linguistics.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7360–7370. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020a. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. Association for Computational Linguistics.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. 2020b. Robust pre-training by adversarial contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020a. Exploiting unsupervised data for emotion recognition

in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4839–4846. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020b. Real-time emotion recognition via attention gated hierarchical memory network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8002–8009. AAAI Press.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial self-supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 443–455. Association for Computational Linguistics.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5669–5679. Association for Computational Linguistics.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021,*

Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 1204–1214. Association for Computational Linguistics.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021b. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022a. Contrast and generation make BART a good dialogue emotion recognizer. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11002–11010. AAAI Press.

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022b. AMOA: global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 7136–7146. International Committee on Computational Linguistics.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Guang Chen. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2543–2556. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

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.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Inf. Fusion*, 64:50–70.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational*

Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6818–6825. AAAI Press.

Manuel López Martín, Antonio Sánchez-Esguevillas, Juan Ignacio Arribas, and Belén Carro. 2022. Supervised contrastive learning over prototype-label embeddings for network intrusion detection. *Inf. Fusion*, 79:200–228.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

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

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13824–13833.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3353–3364.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. DialogXL: all-in-one xlnet for multi-party conversation emotion recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI*

2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 13789–13797. AAAI Press.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1551–1560. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1849–1857.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5197–5206. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 943–952. Computer Vision Foundation / IEEE.

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, SIGdial 2020,*

1st virtual meeting, July 1-3, 2020, pages 186–195. Association for Computational Linguistics.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, volume 12459 of *Lecture Notes in Computer Science*, pages 633–649. Springer.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11595–11603. AAAI Press.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Technical Report*, pages 44–52. AAAI Press.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. 2019a. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 227–238.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019b. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421. ijcai.org.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. CauAIN: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4524–4530. ijcai.org.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 165–176. Association for Computational Linguistics.

Yan Zhou, Longtao Huang, Tao Guo, Jizhong Han, and Songlin Hu. 2019. A span-based joint model for opinion target extraction and target sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5485–5491. ijcai.org.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1571–1582. Association for Computational Linguistics.

Appendix Overview

In this supplementary material, we provide: (i) the related work, (ii) a detailed description of experimental setups, and (iii) detailed results.

A Related Work

A.1 Emotion Recognition in Conversations

Unlike traditional sentiment analysis (Zhou et al., 2019; Wei et al., 2020; Hu et al., 2022c; Li et al., 2022b), context information plays a significant role in identifying the emotion in conversations (Poria et al., 2019b). Existing works usually utilize deep learning techniques to identify the emotion by context modeling and emotion representation learning. These works can be roughly divided into sequence-, graph- and Transformer-based methods.

A.1.1 Sequence-based Methods

Sequence-based methods (Poria et al., 2017; Hazarika et al., 2018b,a; Majumder et al., 2019; Ghosal et al., 2020; Jiao et al., 2020a,b; Hu et al., 2021a; Zhao et al., 2022) generally utilize sequential information in a dialogue to capture different levels of contextual features, i.e., situation, speakers and emotions. For example, Poria et al. (2017) employ an LSTM to capture context-level features from surrounding utterances. Hazarika et al. (2018b,a); Jiao et al. (2020b) use memory networks to capture contextual features. Majumder et al. (2019) use GRUs to capture speaker, context and emotion features. Jiao et al. (2020a) introduce a conversation completion task based on unsupervised data to benefit the ERC task. Ghosal et al. (2020); Zhao et al. (2022) utilize GRUs to fuse commonsense knowledge and capture complex interactions in the dialogue. Hu et al. (2021a) propose a cognitive-inspired network that uses multi-turn reasoning modules to capture implicit emotional clues in conversations. In this paper, we propose a supervised adversarial contrastive learning framework with contextual adversarial training to learn class-spread structured representations for better emotion recognition.

A.1.2 Graph-based Methods

Graph-based methods (Ghosal et al., 2019; Zhang et al., 2019b; Ishiwatari et al., 2020; Shen et al., 2021b; Hu et al., 2021b, 2022b; Bao et al., 2022) usually design a specific graph structure to capture complex dependencies in the conversation. For example, Ghosal et al. (2019); Zhang et al. (2019b);

Shen et al. (2021b) leverage GNNs to capture complex interactions in a conversation. In order to simultaneously consider speaker interactions and sequence information, Ishiwatari et al. (2020) introduce a positional encoding module into RGAT. Hu et al. (2021b, 2022b) respectively design a graph-based fusion method that can simultaneously fuse multimodal knowledge and contextual features.

A.1.3 Transformer-based Methods

Transformer-based methods (Zhong et al., 2019; Wang et al., 2020; Shen et al., 2021a; Zhu et al., 2021; Li et al., 2021a; Lee and Choi, 2021; Lee and Lee, 2022; Li et al., 2022a; Song et al., 2022) usually exploit general knowledge in pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Hu et al., 2022a), and model the conversation by a Transformer-based architecture. For example, Zhong et al. (2019) design a Transformer with graph attention to incorporate commonsense knowledge and contextual features. Wang et al. (2020) use a Transformer with an LSTM-CRF module to learn emotion consistency. Shen et al. (2021a) adopt a modified XLNet to deal with longer context and multi-party structures. Lee and Choi (2021) leverage LSTM and GCN to enhance BERT's ability of context modeling. Yang et al. (2022) apply curriculum learning to deal with the learning problem of difficult samples. Li et al. (2022a) utilize a supervised contrastive term and a response generation task to enhance BART's ability for ERC.

A.2 Contrastive Learning and Adversarial Training

A.2.1 Contrastive Learning

Contrastive learning is a representation learning technique to learn generalized embeddings such that similar data sample pairs are close while dissimilar sample pairs stay far apart (Chopra et al., 2005). Sohn (2016); van den Oord et al. (2018); Bachman et al. (2019); Tian et al. (2020); Hénaff (2020); Chen et al. (2020) utilize self-supervised contrastive learning to learn powerful representations. But these self-supervised techniques are generally limited by the risk of sampling bias and nontrivial data augmentation. Li et al. (2021b) propose prototypical contrastive learning to encode the semantic structure of data into the embedding space. Kim et al. (2020); Jiang et al. (2020b); Fan et al. (2021) add instance-wise adversarial examples during self-supervised contrastive learning to improve model robustness. Recently, Khosla et al. (2020);

Optimization	Attack Strength ϵ (IEMOCAP)								Attack Strength ϵ (MELD)							
Objectives	0	0.125	0.25	0.5	1	2	4	Avg. ($\epsilon > 0$)	0	1	2	4	8	16	32	Avg. ($\epsilon > 0$)
CE	68.17	63.50	60.23	55.20	46.01	35.49	27.83	48.04	65.64	54.78	47.59	37.66	28.71	22.29	18.54	34.93
CE with AT	68.55	63.10	59.98	55.00	48.76	39.95	32.31	49.85	65.69	54.29	48.00	40.59	33.68	28.80	26.09	38.58
SACL	**69.22**	**64.59**	**61.76**	**57.41**	**49.81**	**40.77**	**32.80**	**51.19**	**66.45**	**62.03**	**59.96**	**56.35**	**50.88**	**44.69**	**37.69**	**51.93**
Improve	+0.67	+1.09	+1.53	+2.21	+1.05	+0.82	+0.49	+1.34	+0.76	+7.25	+11.96	+15.76	+17.20	+15.89	+11.60	+13.35

Table 8: Context robustness results against different optimization objectives on IEMOCAP and MELD. We report the robust weighted-F1 scores under different attack strengths.
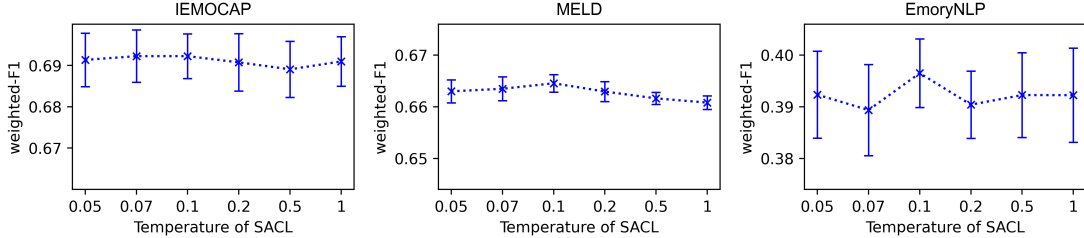

Figure 7: Classification performances of SACL-LSTM against different temperature coefficients on three datasets.

Gunel et al. (2021) use supervised contrastive learning to avoid the above risks and boost performance on downstream tasks by introducing label-level supervised signals. Wang et al. (2021); Martín et al. (2022) use supervised contrastive learning over prototype-label embeddings to learn representations for classification. Lin et al. (2022) employ supervised contrastive learning and CE-based adversarial training to learn domain-adaptive features for low-resource rumor detection. In this paper, we propose a supervised adversarial contrastive learning framework with contextual adversarial training to learn class-spread structured representations for classification on context-dependent data.

A.2.2 Adversarial Training

Adversarial training is a widely used regularization method to improve model robustness by generating adversarial examples with a min-max training recipe (Szegedy et al., 2014). For example, Szegedy et al. (2014) train neural networks on a mixture of adversarial examples and clean examples. Goodfellow et al. (2015) further propose a fast gradient sign method to produce adversarial examples during training. Miyato et al. (2017) extend adversarial and virtual adversarial training to the text domain by applying perturbations to the word embeddings. After that, there are many variants established for supervised/semi-supervised learning (Shafahi et al., 2019; Zhang et al., 2019a; Qin et al., 2019; Jiang et al., 2020a; Zhu et al., 2020).

B Experimental Setups

We report the detailed hyperparameter settings of SACL-LSTM on three datasets in Table 9. The

Hyperparameter	IEMOCAP	MELD	EmoryNLP
Embedding size d_u	1024	1024	1024
Hidden size d_h	128	128	128
Number of LSTM layers	2	1	1
Threshold ξ of Dual-LSTM	2	2	2
Perturbation radius ϵ of CAT	5	5	0.5
Perturbation rate of CAT	1	1	0.1
Norm constraint L_q of CAT	L_2	L_2	L_2
Trade-off weight λ of SACL	0.05	0.1	0.5
Trade-off weight $\lambda^{\text{r-adv}}$ of SACL	0.5	0.05	0.1
Temperature τ and $\tau^{\text{r-adv}}$ of SACL	0.1	0.1	0.1
Number of epochs	100	100	100
Patience	20	20	20
Mini-batch size	2	16	16
Gradient accumulation steps	16	2	2
Learning rate	$1e^{-4}$	$1e^{-4}$	$5e^{-5}$
Weight decay	$2e^{-4}$	$2e^{-4}$	$2e^{-4}$
Dropout	0.2	0.2	0.2
Maximum token length	200	200	200
Flag of class weight	True	False	True
Factor of sample weight	-	1	1

Table 9: Hyperparameter settings of SACL-LSTM on three datasets.

class weights in the CE loss are applied to alleviate the class imbalance issue and are set by their relative ratios in the train and validation sets, except for MELD, which presents a poor effect. For MELD and EmoryNLP, we use focal loss (Lin et al., 2017), a modified version of the CE loss, to balance the weights of easy and hard samples during training.

C Experimental Results

C.1 Results of Context Robustness Evaluation

The detailed results of context robustness evaluation on IEMOCAP and MELD are listed in Table 8.

C.2 Parameter Analysis

Figure 7 illustrates the effect of the temperature parameter in SACL framework on the ERC task.