Self-adaptive Context and Modal-interaction Modeling For Multimodal Emotion Recognition

Haozhe Yang[†], Xianqiang Gao[†], Jianlong Wu[§*]
Tian Gan[†], Ning Ding[‡], Feijun Jiang[‡], Liqiang Nie[§]
[†] Shandong University, [§] Harbin Institute of Technology (Shenzhen), [‡] Alibaba Group
yanghaozhe.sdu@outlook.com, gaoxianqiang@mail.sdu.edu.cn, jlwu1992@pku.edu.cn,
gantian@sdu.edu.cn, {yuji.dn, feijun.jiangfj}@alibaba-inc.com, nieliqiang@gmail.com

Abstract

The multimodal emotion recognition in conversation task aims to predict the emotion label for a given utterance with its context and multiple modalities. Existing approaches achieve good results but also suffer from the following two limitations: 1) lacking modeling of diverse dependency ranges, i.e., long, short, and independent context-specific representations and without consideration of the different recognition difficulty for each utterance; 2) consistent treatment of the contribution for various modalities. To address the above challenges, we propose the Self-adaptive Context and Modal-interaction Modeling (SCMM) framework. We first design the context representation module, which consists of three submodules to model multiple contextual representations. Thereafter, we propose the modal-interaction module, including three interaction submodules to make full use of each modality. Finally, we come up with a self-adaptive path selection module to select an appropriate path in each module and integrate the features to obtain the final representation. Extensive experiments under four settings on three multimodal datasets, including IEMOCAP, MELD, and MOSEI, demonstrate that our proposed method outperforms the state-of-the-art approaches.

1 Introduction

Emotion is a crucial part of human conversation. The emotion recognition in conversation task is to analyze each utterance in a conversation and give the corresponding emotion. This task has recently received more and more attention from researchers in both NLP and multimodal fields because of its potential applications, such as human-computer interaction and opinion mining in social media (Chatterjee et al., 2019; Majumder et al., 2020). Traditional emotion recognition in conversation paradigms is either based on unrelated utterances in a dialogue or a single modality, such as text. However, in many cases, people's emotions are elusive and cannot be delivered well by just one utterance or a single modality. As multimodality is closer to real-world application scenarios, multimodal emotion recognition in conversation is gaining increasing research attention in recent years. To identify emotions more accurately, DialogueRNN (Majumder et al., 2019) first designs an RNN-based model which includes four GRUs to model both intra- and inter-speaker relations. DialogueGCN (Ghosal et al., 2019) then uses a graph neural network to model conversations. Later, MMGCN (Hu et al., 2021) proposes a graph-based method under the additional multimodal setting.

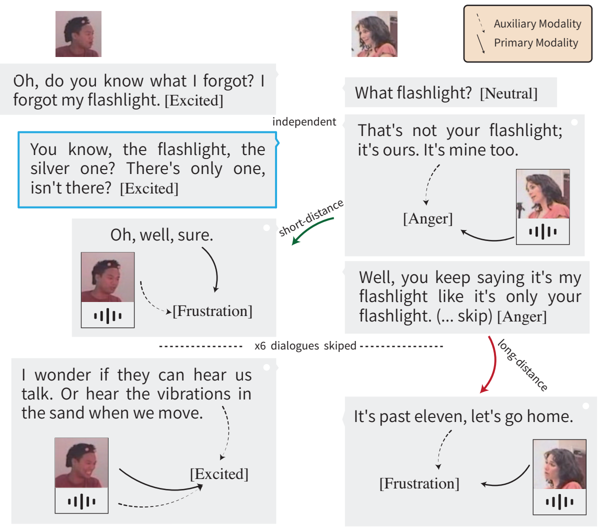

Auxiliary Modality
Primary Modality
Oh, do you know what I forgot? I forgot my flashlight. [Excited]
independent
You know, the flashlight, the silver one? There's only one, isn't there? [Excited]
short-distance
Oh, well, sure.
[Frustration]
x6 dialogues skiped
I wonder if they can hear us talk. Or hear the vibrations in the sand when we move.
[Excited]
What flashlight? [Neutral]
That's not your flashlight; it's ours. It's mine too.
[Anger]
Well, you keep saying it's my flashlight like it's only your flashlight. (... skip) [Anger]
long-distance
It's past eleven, let's go home.
[Frustration]

Figure 1: Motivation of the proposed method. This is an example from the IEMOCAP dataset that contains three different kinds of context dependencies, including long, short, and independent dependency, which is marked by red arrow, green arrow, and blue square, respectively. Besides, the primary modality for the final prediction varies for different samples.

Although pioneer research studies have achieved promising progress, they mainly ignore the varying difficulty of each utterance for the model to recognize and multimodal interaction in conversa-

[*] Jianlong Wu is the corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6267–6281
July 9-14, 2023 ©2023 Association for Computational Linguistics

tion, which leads to the following two limitations. First, existing methods treat all samples equally without considering their specific characteristic or difficulty for recognition. For example, they lack detailed modeling of diverse dependency ranges, i.e., long, short, and independent context-specific representations for each utterance. As illustrated in Figure 1, some utterances in a conversation require a long-range dependency, while others only require a short-range dependency or can determine the emotion on their own. Existing methods do not consider respectively modeling these varying dependency ranges.

Second, current approaches regard the contribution of each modality equally and simply concatenate the features of different modalities. However, the contribution of each modality varies and it is of great importance to investigate the correlation and interaction among different modalities. In particular, Figure 1 illustrates the different contributions among modalities for different utterances, where the primary modality for recognition varies from case to case. We argue the necessity to explore the modality-specific contributions.

Towards the above issues, we propose the Self-adaptive Context and Modal-interaction Modeling (SCMM) method for multimodal emotion recognition. First, to model different ranges of context dependency, we design the context representation module, which consists of three submodules, including global, local, and direct mapping. Second, towards the different contributions of various modalities, we propose the modal-interaction module, which also contains three submodules, including full, partial, and biased interaction, to investigate the correlation among them. Thereafter, faced with multiple outputs from each module, we come up with the self-adaptive path selection strategy to adaptively select an appropriate path to obtain the final representation for each utterance. We also put forward a contrastive learning loss to learn more discriminative representations. Finally, we conduct extensive experiments to validate the effectiveness of our approach.

Our main contributions are four-fold:

- We propose a novel SCMM framework for multimodal emotion recognition in conversation. A new contextual representation module is designed to model different kinds of relation dependency, including long, short, and independent dependency.
- To capture the specific contribution of each modality, we design the modal-interaction module, which consists of three submodules, including full, partial, and biased interactions, to full investigate the correlation among different modalities.
- We come up with the self-adaptive path selection strategy to adaptively select an appropriate path based on module outputs. Moreover, we present a cross-modal contrastive learning loss for discriminative feature learning.
- Extensive experiments on three multimodal emotion recognition datasets, including IEMOCAP, MELD, and MOSEI, demonstrate the superiority of our method. Specifically, on the IEMOCAP dataset under both two different settings, the absolute improvement over state-of-the-art methods is higher than 4.0%.

2 Related Work

2.1 Emotion Recognition in Conversation

Recent years have witnessed growing research interest in Emotion Recognition in Conversation (ERC) due to its wide range of potential applications (Sebe et al., 2005; Yalamanchili et al., 2021). With the development of streaming services, many ERC datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and MOSEI (Bagher Zadeh et al., 2018) provide a new platform for ERC researchers.

To tackle the ERC task, DialogueRNN (Majumder et al., 2019) first proposes an RNN-based model which consists of four GRUs: Global, Speaker, Party, and Emotion, to keep track of the individual and global contextual states in the conversation simultaneously. Following that, DialogueGCN (Ghosal et al., 2019) presents a graph-based model that uses a context window to capture local contextual information. Later, DAGERC (Shen et al., 2021b) applies GNN to construct directed acyclic graphs in conversations and RNN to model local contextual representations. Moreover, COGMEN (Joshi et al., 2022) and MMGCN (Hu et al., 2021) adopt graph-based methods in the same period to model local and global contextual representations, respectively.

Previous work in ERC can be roughly divided into unimodal (Yu et al., 2019; Shen et al., 2021a; Wang et al., 2020) and multimodal approaches (Datcu and Rothkrantz, 2015; Wöllmer et al., 2010). The former uses a single textual modality in experiments, whereas the latter considers acoustic, textual, and visual modalities at the same time. We focus on the multimodal setting.

2.2 Multimodal Fusion

Multimodal fusion aims to make full use of the information in various modalities to improve the recognition results (Atrey et al., 2010; Bramon et al., 2011). This strategy is simple and effective, which has drawn many researchers' attention. For example, in ERC scenarios, DialogueRNN (Majumder et al., 2019) first conducts experiments with single text modality settings but also concatenates multimodal features as an additional experiment. Furthermore, COGMEN (Joshi et al., 2022) follows the setting of concatenating modality in DialogueRNN and designs a GNN model based on this setting. Moreover, MMGCN (Hu et al., 2021) and EmoCaps (Li et al., 2022) concatenate each modality together after passing it through a simple LSTM or linear layer.

However, the multimodal interactions of the existing efforts are still very simple and inevitably lead to suboptimal performance. For example, COGMEN and MMGCN simply concatenate the features of different modalities. We argue that the contribution of different modalities varies and should be treated separately. It is of vital importance to exploit the modal-interaction.

3 Method

3.1 Problem Formulation

In ERC, a conversation is defined as a sequence of utterances $C = \{u_1, u_2, \ldots, u_n\}$, where n is the number of utterances. Each utterance u_i can be labeled by a discrete value y_i, where $y_i \in S$ and S is the emotion labels set. This task aims to predict the emotion label y_i for a given query utterance u_t based on the dialogue context u_1 to u_n and the corresponding speaker identity. Each conversation dataset D contains N dialogues and can be denoted as $D = \{C_j | j = 1, \ldots, N\}$.

In a general multimodal setting, each utterance u_i consists of three modalities, including audio, text, and video, so u_i can be further expressed as $u_i = \{u_i^a, u_i^t, u_i^v\}$, where u_i^a, u_i^t, u_i^v denote the acoustic, textual, and visual features of the i-th utterance with dimension d^a, d^t, d^v, respectively. The whole conversation feature of each modality can be denoted as U^a, U^t, U^v.

3.2 Overview of the Proposed SCMM

As illustrated in Section 1, existing methods do not consider the specific characteristic of diverse dependency ranges for different samples and simply concatenate multimodal features, leading to undesirable results. Therefore, we propose Self-adaptive Context and Modal-interaction Modeling (SCMM) for Multimodal Emotion Recognition. As shown in Figure 2(a), our model first takes the features of each modality as input and obtains the context representation of each modality after passing through the context representation module. Then, each context-represented modality feature will fully interact and complement the information from each other in the modal-interaction module, after which we use self-adaptive path selection module to select appropriate features to get the multimodal representation for final classification.

In the context representation module, we develop three submodules to obtain context representation for utterances with different dependency ranges. First, with the help of the attention mechanism, each utterance can attend to the information of other utterances, so we use a Transformer structure to extract global context representation for a long dependency range. Besides, the GRU structure contains a gate mechanism that can filter out information from long-distance utterances, so we use this unit to obtain the local contextual representation of the utterance for a short dependency range. Finally, for utterances that do not need the assistance of contextual information, we use a linear layer to extract the information. The arrows within each submodule of the context representation module illustrated in left of Figure 2(a) indicate the afflux type of contextual information during the representation process.

For multimodal features, we also consider the difficulty of each utterance for the model to recognize and model it by three modality interaction submodules. For simple utterances, e.g., sentences that contain emotional words, we directly concatenate all modality features together and pass through the linear layer. For slightly complex utterances, we use diverse combinations and interactions among modalities. For more difficult utterances, we take the text modality as the primary modality and others as the auxiliary modalities for interaction. An additional Transformer with a local attention mask is applied to leverage more modality information from adjacent utterances in this phase.

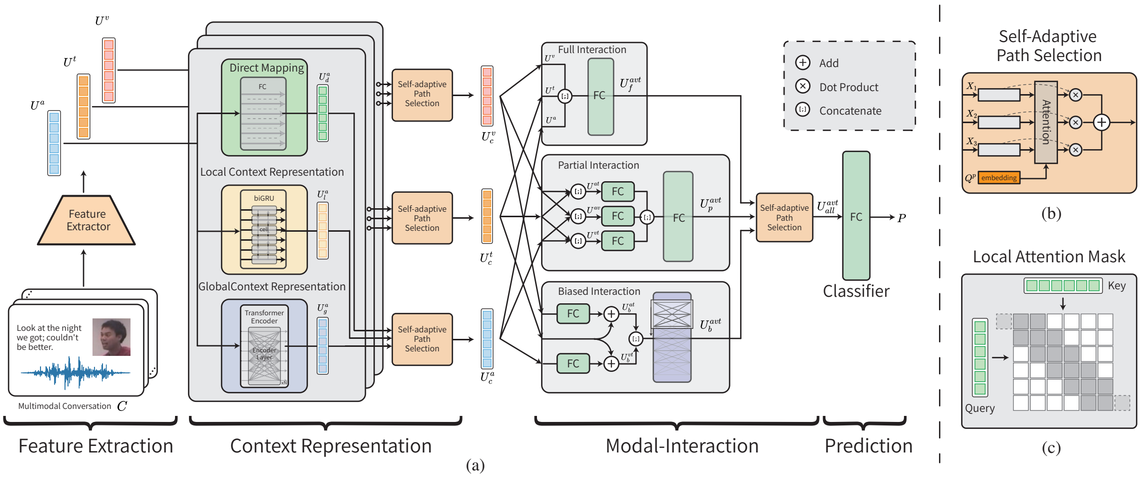


Figure 2: The overview of our proposed SCMM. (a): The overall framework. After feature extraction, the multimodal features go through context representation, modal-interaction, and self-adaptive path selection modules in turn and finally get predictions by the classifier. (b): The internal structure of the self-adaptive path selection module drawn in (a). (c): The diagram of the local attention mask.

3.3 Context Representation Module

Integrating contextual information into the features of utterances is essential, but the demands to establish dependencies between different utterances vary. These dependencies can be summarized into three basic types: long, short, and independent dependency. Based on these different requirements, we design three submodules to consider each case separately.

Global Context Representation: People may discuss several topics in a conversation, and different topics may have different emotional vibes. The current utterance's emotion may be based on another topic raised a relatively long time ago, which is a long-distance emotional dependency relationship. We design the global context representation submodule to model this scenario. With the commonly used attention mechanism, each utterance can attend to other utterances without considering the distance, which ensures effectiveness during long-distance context representation. We use the following multi-head self-attention mechanism to capture global contextual information:

$$\begin{aligned}\mathrm{MultiHead}(Q, K, V) = \\ \mathrm{Concat}(h_1, h_2, \dots, h_n)W^K,\end{aligned} \tag{1}$$

where $Q, K,$ and V are feature matrices, and $Q, K, V \in \mathbb{R}^{n \times d}$. For the self-attention mechanism, Q, K and V are derived from input features with separate linear layers. They will be equally divided into k heads along the feature dimension, the i-th head can be denoted as $Q_i, K_i, V_i \in \mathbb{R}^{n \times \frac{d}{k}}$. $h_i = Attn(Q_i, K_i, V_i)$, and Attn is calculated by Eq. (2) for each head:

$$\mathrm{Attn}(Q_i, K_i, V_i) = \sigma(\frac{Q_i K_i^T}{\sqrt{k}} V_i), \tag{2}$$

where σ denotes the softmax operation.

For dialogue features U^x of different modality, where $x \in \{a, t, v\}$ and $U^x \in \mathbb{R}^{n \times d_x}$, the intermediate representation obtained by $\mathrm{MultiHead}$ is then passed through the commonly used residual concatenation, LayerNorm, and feed-forward layers to obtain the final output U_g^x of this submodule, i.e., U_g^a, U_g^t, and U_g^v.

Local Context Representation: In multi-turn conversations, the emotion of a speaker's utterance may be influenced by adjacent utterances, which is a short-distance emotional dependency that occurs at a local scale. To handle this scenario, we design the local context representation module. The Gated Recurrent Units (GRU) update mechanism ensures that each utterance will integrate contextual information from closer utterances while forgetting information about farther utterances. Therefore, we use a bidirectional GRU network to obtain the local context representation of each utterance. For any modality input U^x, the local context representation feature U_l^x is computed by:

$$U_l^x = \mathrm{Concat}([\overrightarrow{GRU}(U^x), \overleftarrow{GRU}(U^x)]). \tag{3}$$

We denote the features of each modality obtained by this submodule as U_l^a, U_l^t, and U_l^v, respectively.

Direct Mapping: For the utterances that contain enough information on their own, the process of context representation may introduce additional noise. Therefore, we design the direct mapping submodule to directly extract information for each utterance through a linear layer as follows:

$$U_d^x = U^x W_d + b_d. \tag{4}$$

In this submodule, the output features of each modality are U_d^a, U_d^t, and U_d^v, respectively.

3.4 Modal-interaction Module

Given multimodal features U^a, U^t, and U^v, a multimodal interaction module takes these three features as input and outputs a multimodal feature U^{atv}. By effectively exploiting the potential complementarity of information among these modalities, the multimodal features can be more discriminative, allowing the model to perform better than unimodal models. Considering the different difficulties among utterances, we design different interaction submodules to handle simple, more complex, and difficult scenarios, respectively.

Full Interaction: For simple utterances and ideal cases where the three modalities U^a, U^t and U^v complement each other, and each modality contains relatively equal information, we design the full interaction submodule, which concatenates three multimodal features directly and uses a linear layer to extract multimodal feature. We denote it as U_f^{atv} by linear layer and formulate it as follows:

$$U_f^{atv} = \text{Concat}(U^a, U^t, U^v)W_f + b_f. \tag{5}$$

Partial Interaction: For slightly complex utterances, the contribution of different modalities varies due to the lack of key information or the mixing of noise. In this regard, we design the partial interaction submodule to alleviate this problem through diversified modality interactions. Specifically, we combines U^a, U^v and U^t in pairs to obtain U^{at}, U^{vt} and U^{av} features. For example,

$$U^{at} = \text{Concat}(U^a, U^t)W_{at} + b_{at}. \tag{6}$$

Finally, we concatenate all paired features and reduce the dimension by a linear layer. We denote this feature as U_p^{atv}.

Biased Interaction: For more difficult utterances, we design the biased interaction submodule. In previous work, many experiments have shown that textual modality features are critical to the performance of the final model in predicting emotions, which indicates that the textual modality contains the primary information in most cases. Therefore, in this interaction process, we first take the text as the primary modality and others as auxiliary modalities to alleviate the information loss of text. Second, we use a small Transformer with a local attention mask to further leverage more modality information from adjacent utterances.

Specifically, the biased interaction submodule first concatenates U^t together with U^a and U^v respectively to obtain U_b^{at} and U_b^{vt}. These two features will be concatenated after passing through their respective linear layers. Later, a Transformer with a local attention mask is applied to incorporate multimodal information from locally scaled multimodal features.

Take Q, K from the self-attention mechanism, the attention mask can be a binary matrix of dimension $\mathbb{R}^{n \times n}$. $M_{i,j} = 1$ means Q_i can attend to K_j during the attention process. Otherwise, it means not. The operation of masked attention can be formulated as follows:

$$\text{Attn}(Q,K,V,M) = \left[\frac{M \odot \exp\left(QK^T/\sqrt{d_k}\right)}{\sum_i M \odot \exp\left(QK_i^T/\sqrt{d_k}\right)}\right] V, \tag{7}$$

where $\odot$ represents element-wise multiplication.

For the local attention mask of this part, we define the parameters w_p, w_f for length of the dependency context and the binary vector $M_i \in \mathbb{R}^n$, with the value of the j-th element in M_i being:

$$M_{i,j} = \begin{cases} 1, & j - i > w_f \text{ or } i - j > w_p, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Eventually, we obtain the local attention mask $M = [M_1, M_2, ..., M_n]$, where $M \in \mathbb{R}^{n \times n}$. The final multimodal feature U_b^{atv} is obtained after the Transformer with the local attention mask M.

3.5 Self-Adaptive Path Selection

To best take advantage of the outputs of submodules obtained in Sections 3.3 and 3.4, we design the self-adaptive path selection module to adaptively select the most appropriate route and integrate them by groups for the next stage. The path selection process is done in a soft way, like an attention mechanism. As illustrated in Figure 2(b), for a given feature X_1, X_2, X_3 with the same dimension, we first calculate the similarity with these features through a trainable parameter Q^p to get the score of each feature. Then, the normalized score is used as the weight of each feature. We use the softmax operation as the normalized function. Finally, we take the weighted average of these features as the final output, which can be formulated as follows:

$$\begin{aligned}&\mathrm{Select}(X_1, X_2, X_3) = \\ &\quad \sigma(\frac{Q^p[X_1, X_2, X_3]^T}{\sqrt{d^x}})[X_1, X_2, X_3]^T,\end{aligned} \tag{9}$$

where $[\cdot,\cdot]$ is the feature concatenation operation.

In the context representation module, we denote the output of each modality's context representation modality through self-adaptive path selection as U_c^a, U_c^t, U_c^v. In the modal-interaction module, we obtain the feature U_{all}^{atv} as the final multimodal feature by $\mathrm{Select}(U_f^{atv}, U_p^{atv}, U_b^{atv})$.

3.6 Cross-modal Contrastive Learning

We obtain the final prediction by passing U_{all}^{atv} through a linear layer, and the final emotion label $\hat{Y}$ of the input dialogue U can be calculated by softmax (denoted by σ) and $\arg\max$ operations:

$$\begin{aligned}P &= \sigma(U_{all}^{atv} W_2 + b_2), \\ \hat{Y} &= \arg\max(P).\end{aligned} \tag{10}$$

We first define the following classification loss:

$$\mathcal{L}_{cls} = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log p_{i,j}\left[y_{i,j}\right], \tag{11}$$

where N is the number of dialogues, $c(i)$ is the number of utterances in the i-th dialogue, $p_{i,j}$ is the probability distribution of utterance j in the i-th dialogue, and $y_{i,j}$ is the expected class label of utterance j in the i-th dialogue.

In order to improve the discriminability of multimodal features we introduce supervised cross-modal contrastive loss in the modal-interaction module. In this stage, all dialogues within the batch are flattened into utterance feature sequences. For any two features of the same dimension $X_1, X_2 \in \mathbb{R}^{C\times d}$, where C denoting the number of utterances in the current batch, the supervised cross-modal contrastive loss is calculated as:

$$l_i = -\frac{1}{|M_i|}\log\frac{\sum_{j,y_j=y_i}^{C} \exp(\mathrm{sim}(x_{1,i}, x_{2,j}/)\tau)}{\sum_{k,y_k\neq y_i}^{C} \exp(\mathrm{sim}(x_{1,i}, x_{2,k})/\tau)}, \tag{12}$$

where $|M_i|$ denotes the number of samples which have the same emotion label as the i-th sample, τ denotes the temperature defined in the original contrastive loss, and $\mathrm{sim}(x_{1,i}, x_{2,i})$ is used to calculate the cosine similarity of the two vectors. The cross-modal contrastive loss L_{cc} between two feature set X_1, X_2 is calculated by:

$$L_{cc}(X_1, X_2) = \frac{\sum_i^C l_i}{C}. \tag{13}$$

Dataset	Number of dialogues(utterances)		
	train	valid	test
IMOECAP-4	120(3600)		31(943)
IEMOCAP-6	120(5810)		31(1623)
MELD	1152(11098)		280(2610)
MOSEI	2247(16261)	300(1868)	675(4640)

Table 1: Statistics of three datasets under four settings.

We set text as the primary modality and assign the cross-modal contrastive loss to these three interaction submodules to get the following six parts:

$$\begin{aligned}L_{cc} = &L_{cc}(U^a, U^t) + L_{cc}(U^v, U^t) + \\ &L_{cc}(U^{av}, U^{vt}) + L_{cc}(U^{av}, U^{at}) + \\ &L_{cc}(U_b^{vt}, U_b^{at}) + L_{cc}(U_b^{at}, U_b^{vt}).\end{aligned} \tag{14}$$

Then we get the overall training objective:

$$\min_{\theta} \mathcal{L} = \mathcal{L}_{cls} + \beta\mathcal{L}_{cc}, \tag{15}$$

where β is a constant to control the loss weight.

4 Experiments and Results

4.1 Experimental Settings

Dataset

We evaluated our method on three benchmark datasets, including IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and MOSEI (Bagher Zadeh et al., 2018), all of which are multimodal datasets with aligned acoustic, textual, and visual information for each utterance in a conversation. In literature, two IEMOCAP settings are used, one with four emotions (IEMOCAP-4) and one with six emotions (IEMOCAP-6), so there are four benchmarks to be compared. For the train/validation/test splits of the dataset, following previous work, we split IEMOCAP and MOSEI according to the setting in (Joshi et al., 2022), and MELD according to the setting in (Hu et al., 2021). Statistics for these three datasets are summarized in Table 1. For more information, please refer to Appendix A.1.

	IEMOCAP	MELD	MOSEI
Acoustic	100	300	300
Visual	100	600	74
Textual	512	342	35

Table 2: Feature dimensions of each dataset.

Feature Extraction

We extracted uniform features to ensure a fair comparison. For IEMOCAP, audio and video features are obtained in the same way as COGMEN (Joshi et al., 2022), and text features are re-extracted by sBERT. For MELD, audio features (size 300) are extracted by OpenSmile toolkit with IS10 configuration (Schuller et al., 2011), video features (size 600) are extracted by DenseNet (Huang et al., 2017) in the same way as MMGCN (Hu et al., 2021), text features are extracted by sBERT. For MOSEI, audio features (size 640) are extracted using librosa [1] with 640 filter banks, video features (size 35) are extracted by Facets, and text features are extracted by sBERT. We presented the dimensions of the final extracted features for each dataset in Table 2.

Compared Baselines

We compared both unimodal and multimodal methods proposed in the emotion recognition field to verify the effectiveness of our model. For unimodal methods, our model was compared with three baselines, including DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019) and DAG-ERC (Shen et al., 2021b). For multimodal baselines, our model was compared with MMGCN (Hu et al., 2021), COGMEN (Joshi et al., 2022) and EMOCAPs (Li et al., 2022). We reimplemented all these methods under the same experimental settings for fair comparison. The BERT structure in the transformers (Wolf et al., 2020) library is adopted as the Transformer structure used in SCMM, and scipy (Virtanen et al., 2020) is used to calculate the F1-score value. For more information, please refer to Appendix A.2.

Implement Details

Our architecture trained on the IEMOCAP dataset has 304 million parameters and takes around 3 minutes to train for 55 epochs on one 2080Ti GPU. We fixed the random seed for all experiments to ensure the reproducibility of our experiments.

We trained our network using the Adam Optimizer with a learning rate of 1e-4. The length of the dependency context w_f and w_p are set to 5 for IEMOCAP and 2 for MELD and MOSEI. In the biased interaction submodule, the Transformer layers used for IEMOCAP, MELD, and MOSEI are 6, 2, and 2, respectively. β is set to 0.2 for MOSEI, and 1 for other datasets. The above optimal parameters are learned based on the grid-search strategy.

Following previous work (Hazarika et al., 2018; Majumder et al., 2019; Ghosal et al., 2019), we used weighted average F1-score for evaluation.

4.2 Main Results

Table 3 shows the results of our model compared with other models on several multimodal emotion conversation datasets. We have the following observations. On the one hand, our method achieves significant improvement over existing state-of-the-art methods. Specifically, our results are 6.84%, 4.44%, 2.36%, and 1.25% absolutely higher than the second best result on IEMOCAP-6, IEMOCAP-4, MELD, and MOSEI, respectively, demonstrating the superiority of our method SCMM.

On the other hand, by comparing the results of last two lines, we can see that the cross-modal contrastive learning loss can bring consistent improvement on all these datasets, where the average improvement is about 0.8%. The reason is that the proposed contrastive loss can benefit the learning of discriminative features and make the margin between different classes more clear.

4.3 Ablation Study and Analysis

Effect of Submodules

We compared the effects of different context representation submodules and modal-interaction submodules. We divided these submodules into three parts based on their complexity, including the direct mapping with the full interaction, the local context representation with the partial interaction, and the global context representation with the biased interaction. We then tested the effectiveness of these three parts. The results are shown in Table 4, where the absence of different modules (w/o U_d^x and U_f^{avt}, w/o U_l^x and U_p^{avt} and w/o U_g^x and U_b^{avt}) exhibits some performance loss on these datasets. Among them, the absence of the global context representation with the biased interaction submodule causes the largest performance loss on all compared datasets. Moreover, we can see that by removing

[1] https://librosa.org/doc/latest/index.html

Models	IEMOCAP-6							IEMOCAP-4	MELD	MOSEI
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average			
DialogueRNN	36.43	67.34	49.62	59.55	63.93	49.35	54.74	74.11	52.44	48.40
DialogueGCN	**56.85**	72.17	48.47	54.17	74.16	50.86	58.68	75.15	57.08	48.40
DAGERC	50.17	73.25	56.55	56.41	66.28	58.27	60.69	73.38	51.01	48.45
MMGCN	33.18	66.96	56.03	63.90	68.14	58.51	59.29	74.81	56.30	59.92
COGMEN	52.31	73.39	53.55	58.97	71.48	53.85	60.38	77.62	55.43	50.50
EmoCaps	22.22	67.27	46.27	56.99	67.86	56.37	54.78	75.14	55.92	48.40
SCMM (w/o L_{cc})	53.23	**79.42**	**63.63**	**66.84**	75.17	60.11	66.73	80.82	58.79	60.70
SCMM (ours)	45.37	78.76	63.54	66.05	**76.70**	**66.18**	**67.53**	**82.06**	**59.44**	**61.17**

Table 3: F1-score comparison on IEMOCAP, MELD, MOSEI datasets. L_{cc} is the cross-modal contrastive loss.

Methods	IEMOCAP-6	IEMOCAP-4	MOSEI
w/o U^t	48.90	69.48	54.47
w/o U^a	64.29	77.64	61.09
w/o U^v	66.08	80.39	60.09
w/o U^a and U^t	39.49	48.91	53.63
w/o U^v and U^t	50.84	66.37	48.48
w/o U^a and U^v	64.83	77.28	59.82
w/o U_d^x and U_f^{avt}	64.76	80.28	60.83
w/o U_l^x and U_p^{avt}	66.14	79.75	59.93
w/o U_g^x and U_b^{avt}	55.49	72.72	59.86
Ours(w/o L_{cc})	66.73	80.82	60.70
Ours	**67.53**	**82.06**	**61.17**

Table 4: Ablation study of our method.

Methods	IEMOCAP-6	IEMOCAP-4	MOSEI
linear selection	65.66	81.39	61.14
self-adaptive path selection	**67.53**	**82.06**	**61.17**

Table 5: Comparison of experimental results using the self-adaptive path selection module and the linear selection module.

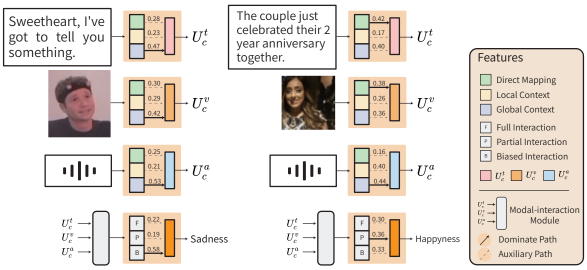


Figure 3: Illustration of weights computed by self-adaptive path selection module for two utterances.

one or two modalities except U^v, especially U^t, the performance will decrease significantly. Above results can also verify that the text is the primary modality for this task.

Effect of Self-adaptive Path Selection

The self-adaptive path selection is designed for the integration of features in different modules. To demonstrate that this module plays a key role in our model, we replaced it with an alternative implementation, where the input features are directly concatenated and then reduced in dimension by a linear layer, which we call the linear selection module. Table 5 shows that replacing our self-adaptive path selection module with the linear selection module leads to performance losses on all datasets, suggesting that the self-adaptive path selection can yield better features.

We also illustrated the weights of each path from several samples to gain deep insights. As shown in Figure 3, in the context representation module, the global context representation submodule is the most important one. In the modal-interaction module, all the cases show that the biased and partial interaction submodules are the most important, which implies that the modal-interaction requires more diverse interaction strategies rather than directly concatenating multimodal features.

Influence of Feature Extractor

For the results in Table 3, we reimplemented all compared baseline methods and used the same extracted features to ensure a fair comparison, which may result in different results than those reported in the paper. To demonstrate the generalization ability of our method, we also conducted additional experiments on IEMOCAP-6 based on the features extracted by COGMEN (Joshi et al., 2022) and EmoCaps (Li et al., 2022). The detailed difference between features can be found in Appendix. The results are shown in Table 6. We can see that our SCMM still achieves much better performance than them under their settings, validating our superiority and robustness.

Methods	F1-score	Methods	F1-score
COGMEN	62.28	EmoCaps	71.16
SCMM (w/o L_{cc})	68.50	SCMM (w/o L_{cc})	73.70
SCMM	**69.08**	SCMM	**75.18**

Table 6: Results comparison under other methods' feature extraction settings on IEMOCAP-6.

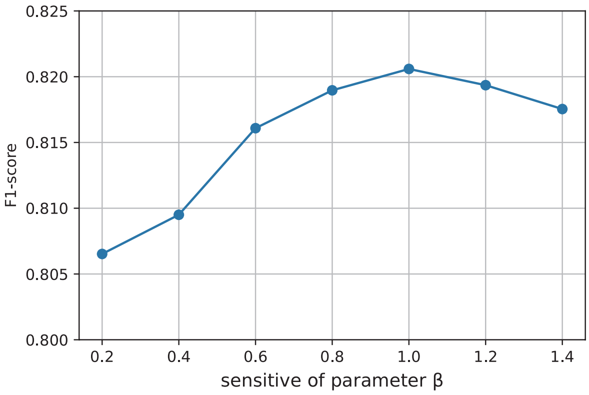


Figure 4: Influence of β for $\mathcal{L}_{cc}$ on IEMOCAP-4.

Parameter Sensitivity Analysis

According to the training objective in Eq. (15), there is mainly one parameter β, which controls the contribution of cross-modal contrastive learning loss. In experiments, we find the optimal value for β by grid searching. We present the results of our method on IEMOCAP-4 with respect to different β in Figure 4. We can observe that our method is relatively stable when β varies in the range of $[0.8, 1.2]$, which show that SCMM is insensitive to this parameter in a certain range.

5 Conclusion

In this paper, for the task of multimodal emotion recognition, we propose the self-adaptive contextual and modal-interaction modeling method. We first come up with the context representation module with global, local modeling and direct mapping to solve the issue of long, short, and independent dependency. Then the modal-interaction consists of full, partial, and bias interactions to fully investigate the correlation and potential complementarity among different modalities. Then we propose the self-adaptive path selection module for better combination and cross-modal contrastive learning loss for discriminative feature learning. Extensive experiments on three datasets under four settings have demonstrated the effectiveness and superiority of our proposed method.

6 Limitations

Our proposed method is an offline system in which the input is a dialogue containing all utterances rather than a single utterance input in chronological order. An online system for emotion recognition can be applied in real-time conference systems or human-computer interaction, so the online system has potential value for future research. Our method can be built into online systems by creating buffer systems such as history windows. However, all the baseline methods in the past are offline systems, such as COGMEN, DialogueRNN, etc. In addition, the form of datasets also leads us to construct an offline system for training and testing. On the other hand, the offline system also has application scenarios such as analyzing emotions of posted videos, opinion mining in social media, etc. Therefore, our method only builds an offline system under the offline experimental setting that can be compared and evaluated.

Besides, the input of our method is feature-based. The original text, audio, and video files will first pass through feature extractors to obtain multimodal features, which may cause information loss and hurt performance. We focus on feature-based training methods because training based on the original files costs a lot. For example, training a video encoder generally requires several V100 GPUs and days of training time. Therefore, we, including the baseline methods we compare, adapt the feature-based training methods. When the cost permits, training based on source files is worth exploring in future work. With feature-based training methods, different baseline methods use feature extractors to obtain features, leading to a lack of fairness in method comparison. In this regard, we reimplemented all open-source methods and compared them using a unified feature file to ensure the fairness of the experimental results. At the same time, we also conducted evaluations with different signature files to verify the generalization of the method.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (grant nos. 62006140, 62176137, and 62236003), in part by the Shandong Provincial Natural Science Foundation (grant no. ZR2020QF106), and in part by the CCF-Alibaba Innovative Research Funds for Young Scholars.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246.

Roger Bramon, Imma Boada, Anton Bardera, Joaquim Rodriguez, Miquel Feixas, Josep Puig, and Mateu Sbert. 2011. Multimodal data fusion based on mutual information. *IEEE Transactions on Visualization and Computer Graphics*, 18:1574–1587.

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.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

Dragos Datcu and Leon JM Rothkrantz. 2015. Semantic audiovisual data fusion for automatic emotion recognition. pages 411–435.

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 Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 154–164.

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

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

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

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

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8968–8979.

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.

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

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53:1062–1087.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2005. Multimodal approaches for emotion recognition: a survey. volume 5670, pages 56–67.

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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1551–1560.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay

Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, pages 261–272.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth S. Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Bhanusree Yalamanchili, Keerthana Dungala, Keerthi Mandapati, Mahitha Pillodi, and Sumasree Reddy Vanga. 2021. Survey on multimodal emotion recognition systems. In *Machine Learning Technologies and Applications*, pages 319–326.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

A Appendix

A.1 Datasets and Feature Extraction

We summarized the statistics for these three datasets in Table 1. All used datasets are commonly used for emotion recognition in the English language. The ids of the data are anonymized by sequential ids or random hash values.

IEMOCAP: IEMOCAP is a multimodal dataset that contains approximately 12 hours of videos for human emotion recognition analysis. Each video consists of a single dyadic dialogue, and every utterance in a conversation is annotated with an emotion label from six categories: happy, sad, neutral, angry, excited, and frustrated. IEMOCAP has two settings, one for four emotion recognition tasks (angry, sad, happy, neutral) and one for six emotion recognition tasks (happy, sad, neutral, angry, excited, and frustrated). We conducted experiments on both of these settings. The IEMOCAP dataset uses the license written by itself, and we have obtained the authorization of The Signal Analysis and Interpretation Laboratory required for accessing and using the IEMOCAP dataset.

MELD: MELD is a large-scale multimodal and multi-speaker emotional dialog dataset collected from the Friends TV series. There are more than 1.4k dialogues in the dataset, and the dialogues are participated by multiple speakers instead of only two. Each utterance in a conversation is annotated with an emotion label from seven categories: anger, disgust, sadness, joy, neutral, surprise, and fear. It uses the GNU (General Public License) v3.0 license.

MOSEI: MOSEI is an emotional recognition dataset made up of 23k sentence utterance video clips taken from YouTube. Specifically, unlike multi-speaker datasets such as IEMOCAP and MELD, MOSEI has only one speaker in a video clip. Each utterance is annotated with an emotion label from six categories: happiness, sadness, disgust, fear, surprise, and anger. CMU-MOSEI also uses a license written by itself, which declaims that the dataset is free for anyone.

We extracted uniform features to ensure a fair comparison. For IEMOCAP, audio and video features are obtained in the same way as COGMEN (Joshi et al., 2022), and text features are re-extracted by sBERT. For MELD, audio features (size 300) are extracted by OpenSmile toolkit with IS10 configuration (Schuller et al., 2011), video features (size 600) are extracted by DenseNet (Huang et al., 2017) in the same way as MMGCN (Hu et al., 2021), text features are extracted by sBERT. For MOSEI, audio features (size 640) are extracted using librosa [2] with 640 filter banks, video features (size 35) are extracted by Facets, and text features are extracted by sBERT.

The distribution of the data used in our evaluation may have some bias. For example, IEMOCAP comes from the performance of some actors, and MELD is obtained from the TV series Friends. In real-world scenarios, conversations may be more complex, such as the position of the camera may be more variable, the types of emotions may be more, the modality of the collected data may be missing, etc. However, all baselines we compared are evaluated on these datasets. In the future, datasets in the wild or collected from natural scenes can be considered to verify the effectiveness of our algorithms.

A.2 Baselines and Implementation

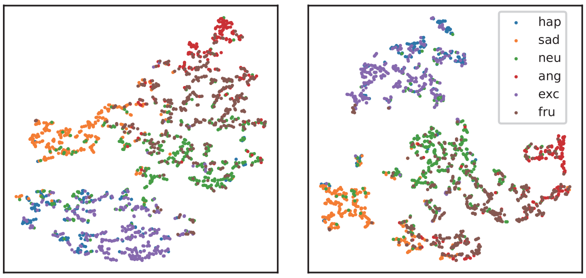
DialogueGCN (Ghosal et al., 2019): it leverages self and inter-speaker dependency based on a graph convolutional network. Each node of the graph represents individual utterance features encoded by bi-LSTM, and the edges between a pair of nodes are constructed relying on the dependency between speakers within a sliding window. Due to only the text modality being used in DialogueGCN, we simply concatenated the features of three modalities for DialogueGCN to make it comparable to SCMM.

DialogueRNN (Majumder et al., 2019): it employs four gated recurrent units(GRU), global GRU, party GRU, and emotion GRU to model the speaker, the context, and the emotion of the preceding utterances. Specifically, the global, party, and speaker GRU update the context, party state, and speaker state, respectively. The emotion GRU is used to model the emotionally relevant representations.

DAG-ERC (Shen et al., 2021b): it models the conversation context through a directed acyclic graph with constraints on speaker identity and positional relations. Furthermore, DAG-ERC gathers contextual information for utterances in a single layer based on a directed acyclic graph neural network.

COGMEN (Joshi et al., 2022): it leverages both local information in a dialogue based on GNN, and the GraphTransformers are used to fuse multiple modalities. However, instead of exploiting the in-

[2] https://librosa.org/doc/latest/index.html



(a) w/o L_{cs} (b) w/ L_{cs}

Figure 5: t-SNE representation of IEMOCAP-6 before and after appling L_{cc}.

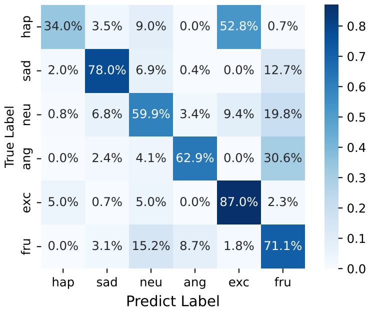


Figure 6: Confusion matrix for IEMOCAP-6.

trinsic connections between features of different modalities, COGMEN simply concatenates them and does not enhance much in multimodal settings.

MMGCN (Hu et al., 2021): it utilizes both multimodal and long-distance contextual information based on a graph convolutional network. In addition, MMGCN constructs graphs in each modality and builds edges between nodes corresponding to the same utterance across multiple modalities. Though good results were achieved on IEMOCAP and MELD, it still treats different modalities in nearly the same way, which somewhat reduces the performance on multimodal tasks.

EmoCaps (Li et al., 2022): it designs a model named Emoformer based on Transformer for feature extraction. After feature extraction, the three modality features are concatenated. Finally, a model based on bi-LSTM layers is applied for emotion prediction.

We used PyTorch to reimplement all these methods and SCMM. The BERT structure in the transformers (Wolf et al., 2020) library is adopted as the Transformer structure used in SCMM, and scipy (Virtanen et al., 2020) is used to calculate the F1-score value. Our architecture trained on the IEMOCAP dataset has 304 million parameters and takes around 3 minutes to train for 55 epochs on one 2080Ti GPU. We fixed the random seed for all experiments to ensure the reproducibility of our experiments.

A.3 Visualization of Contrastive Learning Features

We adopted the t-SNE to visualize feature maps before and after adding the cross-modal contrastive learning loss. As shown in Figure 5, our contrastive learning loss widens the gap among different classes, leading to more discriminative feature representations.

A.4 Error Analysis

After analyzing the dataset, we found that the error predictions of our model mainly came from the error identification of similar emotions. As shown in Figure 6, where most of the error samples in happy are classified as excited and most of the error samples in frustration are classified as anger, etc. These problems also exist in DialogueRNN, COGMEN, and DAGERC. Even though our final results show some improvement compared to previous work, the model still cannot avoid such prediction bias.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6.

☑ A2. Did you discuss any potential risks of your work?
Appendix A

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
No AI writing assistants are used.

B ☑ Did you use or create scientific artifacts?

Section 3.

☑ B1. Did you cite the creators of artifacts you used?
section 4 and Appendix A.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.