MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations

Tao Shi
Tsinghua Shenzhen International Graduate School,
Tsinghua University
shitao21@mails.tsinghua.edu.cn

Shao-Lun Huang*
Tsinghua Shenzhen International Graduate School,
Tsinghua University
shaolun.huang@sz.tsinghua.edu.cn

Abstract

Emotion Recognition in Conversations (ERC) is an increasingly popular task in the Natural Language Processing community, which seeks to achieve accurate emotion classifications of utterances expressed by speakers during a conversation. Most existing approaches focus on modeling speaker and contextual information based on the textual modality, while the complementarity of multimodal information has not been well leveraged, few current methods have sufficiently captured the complex correlations and mapping relationships across different modalities. Furthermore, existing state-of-the-art ERC models have difficulty classifying minority and semantically similar emotion categories. To address these challenges, we propose a novel attention-based correlation-aware multimodal fusion framework named MultiEMO, which effectively integrates multimodal cues by capturing cross-modal mapping relationships across textual, audio and visual modalities based on bidirectional multi-head cross-attention layers. The difficulty of recognizing minority and semantically hard-to-distinguish emotion classes is alleviated by our proposed Sample-Weighted Focal Contrastive (SWFC) loss. Extensive experiments on two benchmark ERC datasets demonstrate that our MultiEMO framework consistently outperforms existing state-of-the-art approaches in all emotion categories on both datasets, the improvements in minority and semantically similar emotions are especially significant.

1 Introduction

Emotion Recognition in Conversations (ERC) is an emerging task in the field of Natural Language Processing (NLP), which aims to identify the emotion of each utterance in a conversation based on textual, audio and visual cues of the speaker. ERC has attracted an enormous amount of attention from both academia and industry, due to its widespread potentials in social media analysis (Chatterjee et al., 2019), health care services (Hu et al., 2021b), empathetic systems (Jiao et al., 2020), and so on.

To solve the problem of ERC, numerous approaches have been proposed. The majority of existing works concentrate on modeling speaker dependencies and conversational contexts (Poria et al., 2017; Hazarika et al., 2018a,c; Majumder et al., 2019; Ghosal et al., 2019, 2020; Shen et al., 2021; Hu et al., 2021a,b; Li et al., 2021a; Joshi et al., 2022; Lee and Lee, 2022), while there still exist several unsolved challenges: (1) **The complementarity of multimodal information has not been well exploited.** Apart from rich information contained in the textual modality, the tone and intonation of the speaker can indicate the intensity of the emotion, facial expressions of interlocutors are also able to explicitly reveal emotional tendencies (Li et al., 2022). Figure 1 shows an example where the complementarity of acoustic and visual signals in addition to the textual modality is essential for an accurate emotion classification. Nevertheless, most existing approaches focus on the textual modality of utterances or simply utilize feature concatenation as the multimodal fusion mechanism (Poria et al., 2017; Hazarika et al., 2018a,c; Majumder et al., 2019; Zhang and Chai, 2021; Li et al., 2022) without modeling the complicated correlations and mapping relationships across textual, audio and visual modalities, which results in an inadequate integration of multimodal cues. (2) **Unsatisfactory performances in minority emotion classes.** Existing benchmark datasets in ERC, such as IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019), suffer from the problem of imbalanced classes. As illustrated in Figure 2, both MELD and IEMOCAP are class-imbalanced, especially in MELD, where the majority class *neutral* takes up a much larger proportion than minority classes *disgust* and *fear*. Current state-of-the-art approaches fail to solve the class imbalance problem and have poor perfor-

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14752–14766

mances in minority emotions. (3) **The difficulty of distinguishing between semantically similar emotions.** It remains to be a challenging task to correctly classify different emotions that are semantically related, such as *disgust* and *anger* in MELD, since they share similar underlying cognitive, affective and physiological features, and tend to be expressed by speakers in similar contexts.

To address the above problems, in this paper, we propose a novel attention-based correlation-aware multimodal fusion framework named MultiEMO. Firstly, unimodal feature extraction and context modeling are performed for each modality, in which we introduce a visual feature extractor named VisExtNet based on a Multi-task Cascaded Convolutional Network (MTCNN) (Zhang et al., 2016) and a VGGFace2 (Cao et al., 2018) pre-trained ResNet-101 (He et al., 2016). VisExtNet accurately captures visual cues of utterance videos by extracting emotion-rich facial expressions of interlocutors without modeling redundant scene-related visual information. Secondly, we propose a multimodal fusion model called MultiAttn to effectively integrate multimodal information based on bidirectional multi-head cross-attention layers (Vaswani et al., 2017), which successfully captures complex cross-modal correlations and mapping relationships across contextualized textual, audio and visual features. Thirdly, in order to mitigate the difficulty of classifying minority and semantically similar emotion classes, enlightened by Focal Contrastive loss (Zhang et al., 2021), a Sample-Weighted Focal Contrastive (SWFC) loss is proposed, in which we assign more focus to hard-to-classify minority classes and make sample pairs with different emotion labels mutually exclusive with each other such that semantically similar emotions can be better distinguished. In addition, we utilize a Soft Hirschfeld-Gebelein-Rényi (Soft-HGR) loss (Wang et al., 2019) to maximize the correlations across multimodal-fused textual, audio and visual feature representations extracted from MultiAttn. Finally, extensive experiments are conducted on two ERC benchmark datasets, MELD and IEMOCAP. Experimental results demonstrate the effectiveness and superiority of our proposed MultiEMO framework compared with existing state-of-the-art approaches, the improvements in minority and semantically similar emotion categories are especially remarkable.

The main contributions of this work can be summarized as follows:

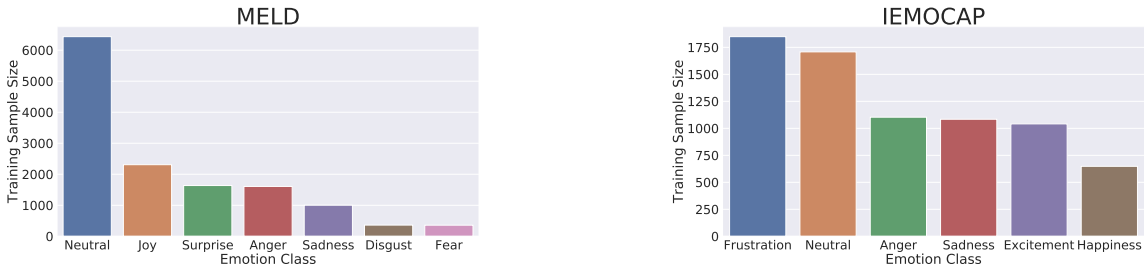
Figure 1: Illustration of the significance of multimodal cues for an accurate prediction, with blue indicating key modalities responsible for the emotion of the utterance.

- We propose a novel visual feature extraction network named VisExtNet, which effectively captures visual cues of interlocutors without modeling redundant scene information.
- We design a multimodal fusion model called MultiAttn based on bidirectional multi-head cross-attention layers, which successfully models the complicated correlations across textual, audio and visual modalities.
- We innovatively introduce a SWFC loss to address the difficulty of classifying minority and semantically similar emotion classes.
- We conduct extensive experiments on MELD and IEMOCAP, results show that our proposed MultiEMO framework achieves state-of-the-art performances on both datasets, the improvements in minority and semantically similar emotions are especially notable.

2 Related Work

2.1 Recurrence-based Models

(Poria et al., 2017) proposes a Long Short-Term Memory (LSTM) based network named BC-LSTM to extract contextual information from dialogues. Interactive Conversational memory Network (ICON) is proposed by (Hazarika et al., 2018b), which models self- and inter-speaker influences based on gated recurrent units (GRUs). (Majumder et al., 2019) introduces a DialogueRNN to model speaker states and contextual information using GRUs. (Lu et al., 2020) proposes a GRU-based Iterative Emotion Interaction Network (IterativeERC), which models emotion interactions by iteratively using predicted emotion labels. (Ma et al., 2022) designs a Multi-View Network (MVN) to model emotion representations of queries from both word- and utterance-level views based on the attention mechanism and bidirectional GRUs.



(a) Class distribution in MELD.

(b) Class distribution in IEMOCAP.

Figure 2: Illustration of the class imbalance problem in MELD and IEMOCAP.

2.2 Graph-based Models

(Ghosal et al., 2019) proposes a Dialogue Graph Convolutional Network (DialogueGCN) to model the conversational context with a directed graph. (Zhang et al., 2019) designs a graph-based model named ConGCN to capture both context- and speaker-sensitive dependencies. (Shen et al., 2021) introduces a Directed Acyclic Neural Network (DAG-ERC) to capture intrinsic structures of conversations using a directed acyclic graph (DAG). A contextualized Graph Neural Network (GNN) based model named COGMEN is proposed by (Joshi et al., 2022), which exploits both local- and global-level information in a conversation.

2.3 Transformer-based Models

(Li et al., 2020) introduces a transformer-based context-sensitive model named HiTrans based on two hierarchical transformers. (Li et al., 2022) designs a transformer-based model called EmoCaps to extract emotional tendencies from multimodal features. CoMPM is introduced by (Lee and Lee, 2022), which consists of a transformer-encoder based context embedding module (CoM) and a pre-trained memory module (PM).

2.4 Multimodal-based Models

Multimodal Fused Graph Convolutional Network (MMGCN) is proposed by (Hu et al., 2021b), which leverages both multimodal information and long-distance contexts. (Li et al., 2021b) introduces a quantum-like framework named QMNN to jointly perform multimodal fusion and conversational context modeling. (Chudasama et al., 2022) designs a multimodal fusion network named M2FNet based on multi-head attention layers to capture cross-modal interactions. A unified framework named UniMSE is proposed by (Hu et al., 2022), in which multimodal representations are fused by injecting acoustic and visual signals into the T5 model.

3 Methodology

3.1 Problem Definition

Given a dialogue which consists of n utterances $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ uttered by speakers $S_{\mathbf{u}_1}, S_{\mathbf{u}_2}, \ldots, S_{\mathbf{u}_n}$, the goal of ERC is to predict the emotion label of each utterance in the dialogue from the pre-defined k-class emotion category set $\mathcal{Y}$. Each utterance has its corresponding textual (t), audio (a) and visual (v) modalities, which can be illustrated as follows:

$$\mathbf{u}_i = \{\mathbf{u}_i^t, \mathbf{u}_i^a, \mathbf{u}_i^v\}, i \in \{1, \ldots, n\} \quad (1)$$

3.2 Model Overview

The overall framework of MultiEMO is illustrated in Figure 3, which is made up of four key components: Unimodal Feature Extraction, Context Modeling, Multimodal Fusion and Emotion Classification. In the subsequent subsections, we discuss each module in detail.

3.3 Unimodal Feature Extraction and Context Modeling

3.3.1 Textual Modality

Existing research often adopts two different paradigms to extract contextualized textual features: (1) **Two-stage paradigm** (Li et al., 2020; Chudasama et al., 2022): Text sequences are first fed into a pre-trained language model to learn utterance-level local textual representations and then to another transformer to generate dialogue-level global textual features by incorporating contextual information in the conversation. (2) **One-stage paradigm** (Kim and Vossen, 2021; Lee and Lee, 2022): Local utterance-level information and global dialogue-level conversational contexts are jointly captured through fine-tuning a single pre-trained language model. We have explored both approaches and experimental results demonstrate

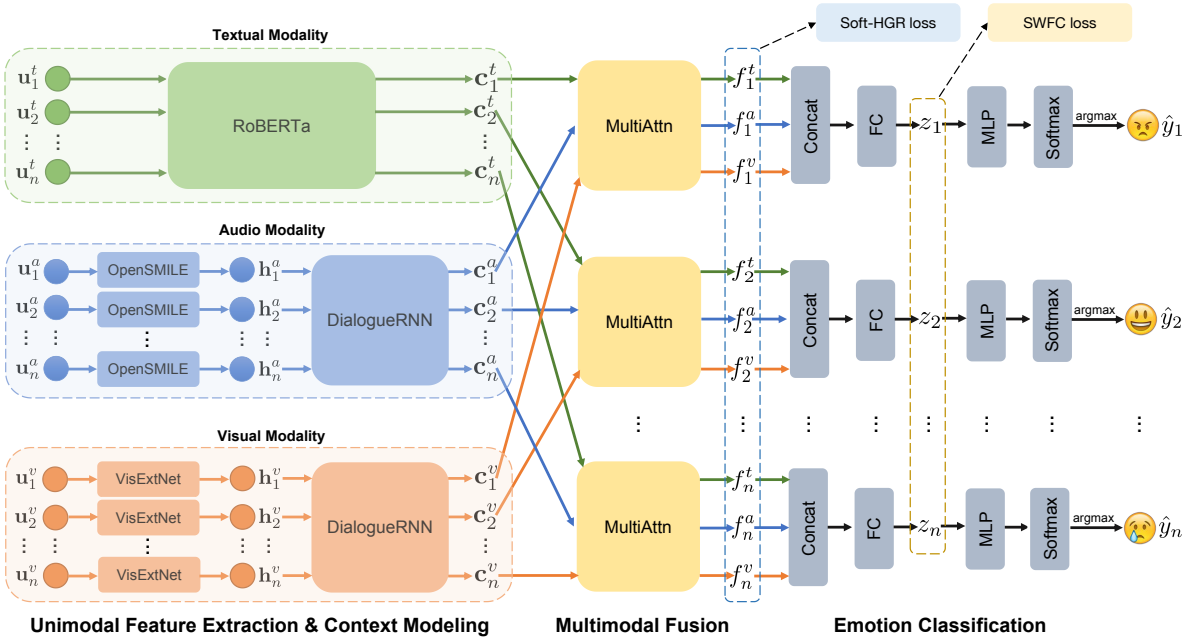


Figure 3: Illustration of the overall framework of MultiEMO, which consists of four key components: Unimodal Feature Extraction, Context Modeling, Multimodal Fusion and Emotion Classification. FC stands for Fully-Connected layer, MLP represents Multilayer Perceptron.

that the one-stage paradigm slightly outperforms the two-stage paradigm. For the sake of computational efficiency, we adopt the one-stage paradigm.

To be specific, following (Kim and Vossen, 2021), each textual utterance $\mathbf{u}_i^t$ is prefixed with the speaker name of the utterance $S_{\mathbf{u}_i}$, such that speaker information can be effectively encoded. Then, the input sequence for the i-th utterance is composed of three segments to incorporate contextual information: preceding contextual utterances $\{\mathbf{u}_1^t, \ldots, \mathbf{u}_{i-1}^t\}$, current utterance $\mathbf{u}_i^t$, and succeeding contextual utterances $\{\mathbf{u}_{i+1}^t, \ldots, \mathbf{u}_n^t\}$. These three segments are concatenated and separated by [SEP] before being fed into a pre-trained RoBERTa model and a subsequent fully-connected layer, with the embedding of the first hidden state [CLS] utilized as the learned contextualized 256-dimensional textual representation $\mathbf{c}_i^t$ for $\mathbf{u}_i^t$.

3.3.2 Audio Modality

Audio Feature Extraction: We follow (Majumder et al., 2019) and use a OpenSMILE (Eyben et al., 2010) to extract a 6373-dimensional feature representation for each utterance audio, then a fully-connected layer is adopted to obtain a 512-dimensional feature $\mathbf{h}_i^a$ for each input audio $\mathbf{u}_i^a$.

Audio Context Modeling: After unimodal audio feature extraction, we employ a DialogueRNN (Majumder et al., 2019) to capture a contextualized audio feature $\mathbf{c}_i^a$ with 256 dimensions for each audio clip. The speaker-modeling nature of DialogueRNN makes it effective in integrating audio cues from different speakers (Li et al., 2022).

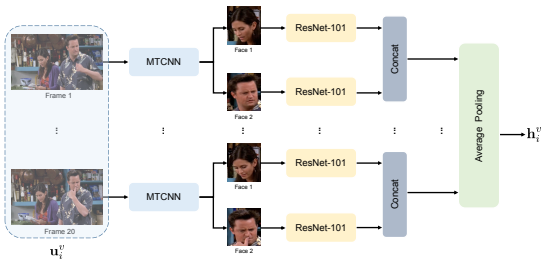


Figure 4: Illustration of the architecture of VisExtNet, which is based on a MTCNN and a VGGFace2 pre-trained ResNet-101.

3.3.3 Visual Modality

Visual Feature Extraction: Most existing works (Hazarika et al., 2018a,c; Majumder et al., 2019; Zhang and Chai, 2021; Li et al., 2022) utilize a 3D-CNN (Tran et al., 2015) to capture visual features from video clips. Recently, (Chudasama et al., 2022) proposes a dual network based on a Multi-task Cascaded Convolutional Network (MTCNN) (Zhang et al., 2016), which demonstrates to be effective. Both approaches encode not only facial expressions of interlocutors but also scene-related information for each utterance clip. However, we

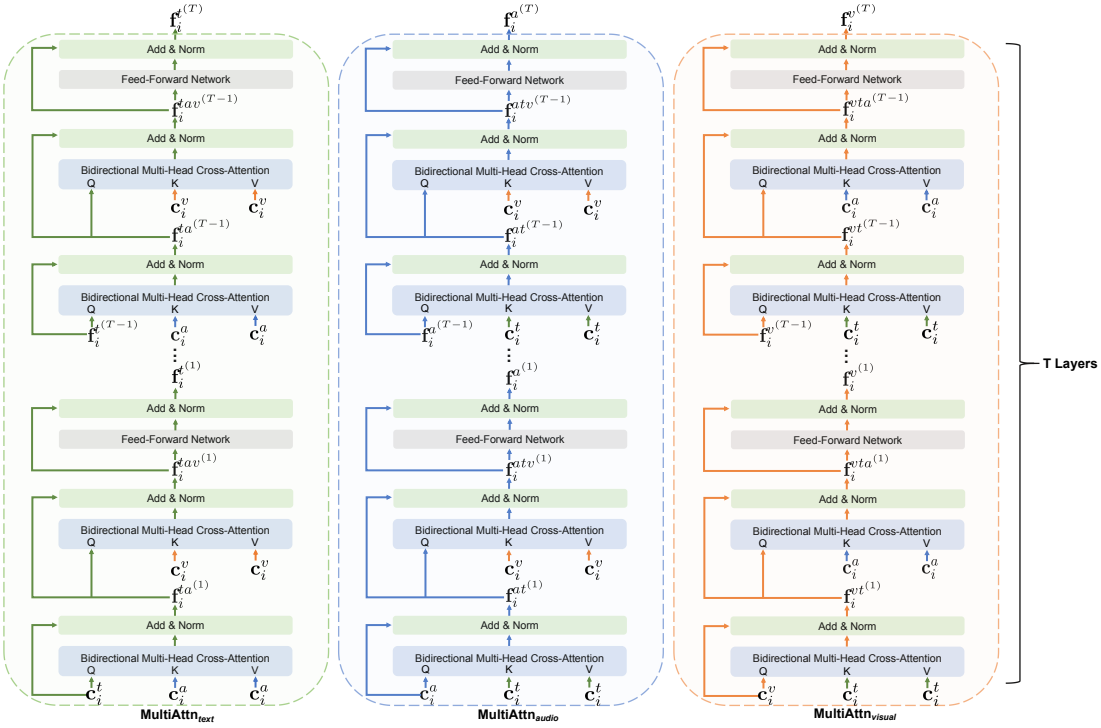


Figure 5: Illustration of the architecture of MultiAttn, which consists of three components: $\text{MultiAttn}_{\text{text}}$, $\text{MultiAttn}_{\text{audio}}$ and $\text{MultiAttn}_{\text{visual}}$, each of which aims to integrate one modality with complementary information from the other two modalities through stacked bidirectional multi-head cross-attention layers.

argue that these two approaches are flawed because visual surrounding information is redundant. Firstly, there are no explicit correlations between scene information and the emotion of the speaker, because dialogues that happen in the same scene do not tend to share similar emotional tendencies. To illustrate, a large proportion of conversations in MELD take place at home, but the emotions of these conversations vary significantly. In addition, the scene normally remains unchanged throughout the conversation. Therefore, capturing scene-related visual information for each utterance is unnecessary and may lead to a wrong understanding of the speaker's actual emotional tendency due to the influence of irrelevant scene information.

To address this problem, we propose a novel visual feature extractor named VisExtNet, which is made up of a MTCNN and a ResNet-101 (He et al., 2016) pre-trained on VGGFace2 (Cao et al., 2018). The architecture of VisExtNet is shown in Figure 4. VisExtNet aims to effectively capture visual cues by integrating facial expressions of interlocutors from multiple frames without encoding redundant scene-related information.

For an utterance video $\mathbf{u}_i^v$, visual feature extraction is performed on 20 frames of the utterance clip, with each frame selected using a step of $\frac{\text{number of frames}}{20}$. Specifically, each frame is first sent into a MTCNN to accurately detect the faces of all interlocutors present in the scene at that frame, each detected face is then passed through a VGGFace2 pretrained ResNet-101 to extract a emotion-rich visual feature vector. The concatenation of facial expression features from all participants is regarded as the visual representation of that frame. The same process is repeated for each of the 20 frames, after which the output features of all frames are average pooled over the frame axis to obtain a 1000-dimensional visual feature vector $\mathbf{h}_i^v$.

Visual Context Modeling: Similar to audio context modeling, after visual feature extraction, we utilize another DialogueRNN to learn a 256-dimensional contextualized visual representation $\mathbf{c}_i^v$ for each video clip.

3.4 Multimodal Fusion

Existing literature fails to effectively integrate multimodal information, the complex correlations and mapping relationships across multiple modalities have not been well captured. To tackle this issue, inspired by (Chudasama et al., 2022), we propose a novel multimodal fusion network named MultiAttn based on the bidirectional multi-head cross-attention mechanism (Vaswani et al., 2017),

in which querys are generated from one modality while keys and values come from a different modality, and both preceding and succeeding contexts are leveraged when calculating attention distributions. The architecture of MultiAttn is shown in Figure 5.

MultiAttn is made up of three components: $\text{MultiAttn}_{\text{text}}$, $\text{MultiAttn}_{\text{audio}}$ and $\text{MultiAttn}_{\text{visual}}$, each of which aims to integrate one modality with complementary information from the other two modalities. As illustrated in Figure 5, $\text{MultiAttn}_{\text{text}}$, $\text{MultiAttn}_{\text{audio}}$ and $\text{MultiAttn}_{\text{visual}}$ share the same building blocks and only differ in terms of input Query, Key and Value. Thus, for the sake of brevity, we use $\text{MultiAttn}_{\text{text}}$ to illustrate how multimodal fusion works. $\text{MultiAttn}_{\text{text}}$ effectively incorporates the textual modality with audio and visual cues through a three-stage approach: (1) $\text{MultiAttn}_{\text{text}}$ first learns cross-modal correlations and mapping relationships between textual and audio modalities by treating the textual modality as Query and the audio modality as Key and Value for the bidirectional multi-head cross-attention operation; (2) The learned output from the first stage is then utilized as the new Query while the visual modality is regarded as Key and Value for another bidirectional multi-head cross-attention layer to fuse the textual modality with visual cues; (3) Finally, a feed-forward network consisting of two fully-connected layers with a Rectified Linear Unit (ReLU) is adopted, which operates as a key-value memory (Geva et al., 2021). In addition, we employ residual connection and layer normalization over the output of each stage to facilitate the training process. To construct a deeper and more powerful network, we utilize T layers of $\text{MultiAttn}_{\text{text}}$, $\text{MultiAttn}_{\text{audio}}$ and $\text{MultiAttn}_{\text{visual}}$ as the full model architecture of MultiAttn, where the output of each layer is fed into the next layer as the new Query.

Given the Queries of all utterances $\mathbf{F}^{t^{(j-1)}} = [\mathbf{f}_1^{t^{(j-1)}}, \ldots, \mathbf{f}_n^{t^{(j-1)}}]^{\mathrm{T}}$ learned from layer $j-1$, audio and visual features $\mathbf{C}^a = [\mathbf{c}_1^a, \ldots, \mathbf{c}_n^a]^{\mathrm{T}}$ and $\mathbf{C}^v = [\mathbf{c}_1^v, \ldots, \mathbf{c}_n^v]^{\mathrm{T}}$, the calculation of $\text{MultiAttn}_{\text{text}}$ at layer j is illustrated as follows:

$$[\mathbf{Q}_h^{ta^{(j)}}, \mathbf{K}_h^{ta^{(j)}}, \mathbf{V}_h^{ta^{(j)}}] = [\mathbf{F}^{t^{(j-1)}}\mathbf{W}^{\mathbf{Q}_h^{ta^{(j)}}}, \mathbf{C}^a\mathbf{W}^{\mathbf{K}_h^{ta^{(j)}}}, \mathbf{C}^a\mathbf{W}^{\mathbf{V}_h^{ta^{(j)}}}], h \in \{1, \ldots, H\} \quad (2)$$

$$\mathbf{A}_h^{ta^{(j)}} = \text{Softmax}(\frac{\mathbf{Q}_h^{ta^{(j)}}\mathbf{K}_h^{ta^{(j)\mathrm{T}}}}{\sqrt{d_{\mathbf{K}_h^{ta^{(j)}}}}})\mathbf{V}_h^{ta^{(j)}}, \quad h \in \{1, \ldots, H\} \quad (3)$$

$$\mathbf{MH}^{ta^{(j)}} = \text{Cat}(\mathbf{A}_1^{ta^{(j)}}, \ldots, \mathbf{A}_H^{ta^{(j)}})\mathbf{W}^{O^{ta^{(j)}}} \quad (4)$$

$$\mathbf{F}^{ta^{(j)}} = \text{LayerNorm}(\mathbf{F}^{t^{(j-1)}} + \mathbf{MH}^{ta^{(j)}}) \quad (5)$$

$$[\mathbf{Q}_h^{tav^{(j)}}, \mathbf{K}_h^{tav^{(j)}}, \mathbf{V}_h^{tav^{(j)}}] = [\mathbf{F}^{ta^{(j)}}\mathbf{W}^{\mathbf{Q}_h^{tav^{(j)}}}, \mathbf{C}^v\mathbf{W}^{\mathbf{K}_h^{tav^{(j)}}}, \mathbf{C}^v\mathbf{W}^{\mathbf{V}_h^{tav^{(j)}}}], h \in \{1, \ldots, H\} \quad (6)$$

$$\mathbf{A}_h^{tav^{(j)}} = \text{Softmax}(\frac{\mathbf{Q}_h^{tav^{(j)}}\mathbf{K}_h^{tav^{(j)\mathrm{T}}}}{\sqrt{d_{\mathbf{K}_h^{tav^{(j)}}}}})\mathbf{V}_h^{tav^{(j)}}, \quad h \in \{1, \ldots, H\} \quad (7)$$

$$\mathbf{MH}^{tav^{(j)}} = \text{Cat}(\mathbf{A}_1^{tav^{(j)}}, \ldots, \mathbf{A}_H^{tav^{(j)}})\mathbf{W}^{O^{tav^{(j)}}} \quad (8)$$

$$\mathbf{F}^{tav^{(j)}} = \text{LayerNorm}(\mathbf{F}^{ta^{(j)}} + \mathbf{MH}^{tav^{(j)}}) \quad (9)$$

$$\mathbf{FFN}_1^{t^{(j)}} = \max(0, \mathbf{F}^{tav^{(j)}}\mathbf{W}^{\mathbf{FFN}_1^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_1^{t^{(j)}}}) \quad (10)$$

$$\mathbf{FFN}_2^{t^{(j)}} = \mathbf{FFN}_1^{t^{(j)}}\mathbf{W}^{\mathbf{FFN}_2^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_2^{t^{(j)}}} \quad (11)$$

$$\mathbf{F}^{t^{(j)}} = \text{LayerNorm}(\mathbf{F}^{tav^{(j)}} + \mathbf{FFN}_2^{t^{(j)}}) \quad (12)$$

Where $\mathbf{W}^{\mathbf{Q}_h^{ta^{(j)}}}$, $\mathbf{W}^{\mathbf{K}_h^{ta^{(j)}}}$, $\mathbf{W}^{\mathbf{V}_h^{ta^{(j)}}}$, $\mathbf{W}^{O^{ta^{(j)}}}$, $\mathbf{W}^{\mathbf{Q}_h^{tav^{(j)}}}$, $\mathbf{W}^{\mathbf{K}_h^{tav^{(j)}}}$, $\mathbf{W}^{\mathbf{V}_h^{tav^{(j)}}}$, $\mathbf{W}^{O^{tav^{(j)}}}$ ($h \in \{1, \ldots, H\}$), $\mathbf{W}^{\mathbf{FFN}_1^{t^{(j)}}}$, $\mathbf{W}^{\mathbf{FFN}_2^{t^{(j)}}}$ are projection matrices, $\mathbf{b}_{\mathbf{FFN}_1^{t^{(j)}}}$ and $\mathbf{b}_{\mathbf{FFN}_2^{t^{(j)}}}$ are bias parameters, H is the number of attention heads, Cat stands for concatenation.

3.5 Emotion Classification

After multimodal fusion, the learned multimodal-fused textual, audio and visual feature representations $\mathbf{f}_i^t$, $\mathbf{f}_i^a$ and $\mathbf{f}_i^v$ are concatenated and then sent into a fully-connected layer and a subsequent 2-layer Multilayer Perceptron (MLP) with a ReLU. Finally, a Softmax layer is utilized to compute a probability distribution over the emotion category set, where the emotion label with the highest probability is chosen as the prediction $\hat{y}_i$ for the i-th utterance. The calculation is illustrated as follows:

$$\mathbf{f}_i = \mathbf{f}_i^t \oplus \mathbf{f}_i^a \oplus \mathbf{f}_i^v \quad (13)$$

$$\mathbf{z}_i = \mathbf{W}^z\mathbf{f}_i + \mathbf{b}_z \quad (14)$$

$$\mathbf{l}_i = \max(0, \mathbf{W}^l\mathbf{z}_i + \mathbf{b}_l) \quad (15)$$

$$\mathbf{p}_i = \text{Softmax}(\mathbf{W}^{smax}\mathbf{l}_i + \mathbf{b}_{smax}) \quad (16)$$

$$\hat{y}_i = \underset{t}{\text{argmax}}(\mathbf{p}_i[t]) \quad (17)$$

Where $\oplus$ denotes concatenation, $\mathbf{W}^z$, $\mathbf{W}^l$ and $\mathbf{W}^{smax}$ are weight matrices, $\mathbf{b}_z$, $\mathbf{b}_l$ and $\mathbf{b}_{smax}$ are bias parameters.

Models	IEMOCAP						
	Happiness	Sadness	Neutral	Anger	Excitement	Frustration	**Weighted-F1**
BC-LSTM	34.43	60.87	51.81	56.73	57.95	58.92	54.95
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	62.75
DialogueGCN	51.87	76.76	56.76	62.26	72.71	58.04	63.16
IterativeERC	53.17	77.19	61.31	61.45	69.23	60.92	64.37
QMNN	39.71	68.30	55.29	62.58	66.71	62.19	59.88
MMGCN	42.34	78.67	61.73	69.00	74.33	62.32	66.22
MVN	55.75	73.30	61.88	65.96	69.50	64.21	65.44
UniMSE	-	-	-	-	-	-	70.66
MultiEMO$_{\text{w/o VisExtNet}}$	65.06	84.80	66.13	67.98	76.16	69.66	71.72
MultiEMO$_{\text{w/o MultiAttn}}$	55.18	78.29	62.06	63.84	73.11	63.98	66.57
MultiEMO$_{\text{w/o SWFC loss}}$	59.88	83.96	66.57	67.03	75.35	70.04	71.08
MultiEMO	**65.77**	**85.49**	**67.08**	**69.88**	**77.31**	**70.98**	**72.84**

Table 1: Experimental results on IEMOCAP. The best results are highlighted in bold. "-" means that the results are unavailable from the original paper.

3.6 Training Objectives

Given a batch of N samples consisting of M dialogues, where the i-th dialogue contains $C(i)$ utterances, training objectives are defined as follows:

SWFC Loss: To alleviate the difficulty of classifying minority and semantically similar emotions, we propose a novel loss function named Sample-Weighted Focal Contrastive (SWFC) loss based on the Focal Contrastive loss (Zhang et al., 2021) loss by introducing a sample-weight term and a focusing parameter, in which we assign more importance to hard-to-classify minority classes during the training phase, and make sample pairs with different emotion labels mutually exclusive with each other to maximize inter-class distances, such that semantically similar emotions can be better distinguished. The SWFC loss is defined as follows:

$$s_{j,g}^{(i)} = \frac{\exp\left(\mathbf{z}_{i,j}^{\mathrm{T}}\mathbf{z}_{i,g}/\tau\right)}{\sum_{\mathbf{z}_{i,s}\in A_{i,j}}\exp\left(\mathbf{z}_{i,j}^{\mathrm{T}}\mathbf{z}_{i,s}/\tau\right)} \tag{18}$$

$$L_{\text{SWFC}} = -\sum_{i=1}^{M}\sum_{j=1}^{C(i)}\left(\frac{N}{n_{y_{i,j}}}\right)^{\alpha}\frac{1}{|R_{i,j}|}\sum_{\mathbf{z}_{i,g}\in R_{i,j}} (1-s_{j,g}^{(i)})^{\gamma}\log s_{j,g}^{(i)} \tag{19}$$

Where $\mathbf{z}_{i,j}$ is the output of the fully-connected layer (Equation 14) for utterance j in dialogue i, $A_{i,j}$ is the set of features in the batch other than $\mathbf{z}_{i,j}$, $y_{i,j}$ is the label of utterance j in dialogue i, $R_{i,j} = \{\mathbf{z}_{i,g} \in A_{i,j} | y_{i,g} = y_{i,j}\}$ is the set of positive features that share the same label as $\mathbf{z}_{i,j}$, $n_{y_{i,j}}$ is the count of label $y_{i,j}$ in the batch, α is a sample-weight parameter that controls the degree of focus on minority classes, τ is a temperature parameter that controls the strength of penalties on negative samples, γ is a focusing parameter which forces the model to focus on hard-to-classify examples.

Soft-HGR Loss: We utilize a Soft Hirschfeld-Gebelein-Rényi (Soft-HGR) loss (Wang et al., 2019) to maximize the correlations across multimodal-fused textual, audio and visual features extracted from MultiAttn. The Soft-HGR loss is defined as follows:

$$\begin{aligned} L_{\text{Soft-HGR}} = &-\sum_{\mathbf{Q}\neq\mathbf{V},\mathbf{Q},\mathbf{V}\in F}(\mathbb{E}[\mathbf{Q}^{\mathrm{T}}\mathbf{V}] \\ &-\frac{1}{2}\text{tr}(\text{cov}(\mathbf{Q})\text{cov}(\mathbf{V}))) \\ &\text{s.t.}\quad \mathbb{E}[\mathbf{Q}] = 0,\ \forall\mathbf{Q}\in F. \end{aligned} \tag{20}$$

Where $F = \{\mathbf{F}^t, \mathbf{F}^a, \mathbf{F}^v\}$, $\mathbf{F}^t = [\mathbf{f}_1^t, \ldots, \mathbf{f}_N^t]^{\mathrm{T}}$, $\mathbf{F}^a = [\mathbf{f}_1^a, \ldots, \mathbf{f}_N^a]^{\mathrm{T}}$, $\mathbf{F}^v = [\mathbf{f}_1^v, \ldots, \mathbf{f}_N^v]^{\mathrm{T}}$. Expectations and covariances are approximated through sample means and sample covariances.

Cross-Entropy Loss: In addition, we adopt a Cross-entropy loss to measure the difference between predicted probabilities and true labels:

$$L_{\text{CE}} = -\sum_{i=1}^{M}\sum_{j=1}^{C(i)}\log\mathbf{p}_{i,j}[y_{i,j}] \tag{21}$$

Where $\mathbf{p}_{i,j}$ is the probability distribution over the emotion classes for utterance j in dialogue i, $y_{i,j}$ is the ground-truth label of utterance j in dialogue i.

Full Loss Function: A linear combination of SWFC loss, Soft-HGR loss and Cross-entropy loss is leveraged as the full loss function:

$$\begin{aligned} L_{\text{Train}} = &\frac{1}{N}(\mu_1 L_{\text{SWFC}} + \mu_2 L_{\text{Soft-HGR}} \\ &+ (1-\mu_1-\mu_2)L_{\text{CE}}) + \lambda||\theta||_2^2,\ \ \mu_1,\mu_2\in[0,1] \end{aligned} \tag{22}$$

Models	MELD							
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	**Weighted-F1**
BC-LSTM	73.80	47.70	5.40	25.10	51.30	5.20	38.40	55.90
DialogueRNN	76.23	49.59	0.00	26.33	54.55	0.81	46.76	58.73
DialogueGCN	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
IterativeERC	77.52	53.65	3.31	23.62	56.63	19.38	48.88	60.72
QMNN	77.00	49.76	0.00	16.50	52.08	0.00	43.17	58.00
MMGCN	-	-	-	-	-	-	-	58.65
MVN	76.65	53.18	11.70	21.82	53.62	21.86	42.55	59.03
UniMSE	-	-	-	-	-	-	-	65.51
MultiEMO$_{\text{w/o VisExtNet}}$	79.16	58.22	24.80	37.61	60.65	31.73	52.08	64.89
MultiEMO$_{\text{w/o MultiAttn}}$	77.72	54.05	21.76	33.10	58.28	24.80	49.98	62.50
MultiEMO$_{\text{w/o SWFC loss}}$	79.51	56.54	20.59	32.96	58.52	25.81	51.23	63.83
MultiEMO	**79.95**	**60.98**	**29.67**	**41.51**	**62.82**	**36.75**	**54.41**	**66.74**

Table 2: Experimental results on MELD. The best results are highlighted in bold. "-" means that the results are unavailable from the original paper.

Modality	IEMOCAP	MELD
Text	64.48	61.23
Audio	38.89	33.55
Visual	35.37	33.16
Text + Audio	69.18	64.21
Text + Visual	67.86	63.78
Text + Audio + Visual	**72.84**	**66.74**

Table 3: Experimental results of MultiEMO with different modality settings on IEMOCAP and MELD.

Where μ_1 and μ_2 are tunable hyperparameters, λ is the L_2 regularization weight, θ is the set of all trainable parameters.

4 Experimental Settings

4.1 Datasets

IEMOCAP (Busso et al., 2008): IEMOCAP contains approximately 12 hours of videos of dyadic conversations, which are segmented into 7433 utterances and 151 dialogues. Each utterance is annotated with one of six emotion labels: happiness, sadness, neutral, anger, excitement and frustration.
MELD (Poria et al., 2019): MELD is a multi-party dataset with 13708 utterances and 1433 dialogues from the TV series *Friends*. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness and surprise.

4.2 Baseline Methods

BC-LSTM (Poria et al., 2017): BC-LSTM models conversational contexts through bidirectional LSTMs without differentiating different speakers.
DialogueRNN (Majumder et al., 2019): DialogueRNN models contextual information and speaker states through three distinct GRUs.
DialogueGCN (Ghosal et al., 2019): DialogueGCN captures the context by modeling conversations using a directed graph.
IterativeERC (Lu et al., 2020): IterativeERC iteratively uses predicted emotion labels instead of gold emotion labels to model emotion interactions.
QMNN (Li et al., 2021b): QMNN captures conversational contexts and conducts multimodal fusion from a novel quantum perspective.
MMGCN (Hu et al., 2021b): MMGCN models long-distance conversational contexts by leveraging Graph Convolutional Networks (GCNs).
MVN (Ma et al., 2022): MVN effectively captures emotion representations of queries from both word- and utterance-level views.
UniMSE (Hu et al., 2022): UniMSE leverages the similarities and complementaries between emotions to achieve better predictions.

4.3 Implementation Details

Modality Setting: We utilize textual, audio and visual modalities of utterances to conduct experiments on both MELD and IEMOCAP.
Hyperparameter Settings: (1) Dataset-specific settings: Since MELD is significantly more class-imbalanced than IEMOCAP, the batch size is designed to be 64 on IEMOCAP and 100 on MELD. (2) Dataset-generic settings: The number of training epochs is 100, the optimizer is Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, the learning rate is initialized with 0.0001 and decays by 0.95 after every 10 epochs, the L_2 regularization weight λ is 0.00001. To avoid overfitting, we apply Dropout (Srivastava et al., 2014) layers with

a dropout rate of 0.1. (3) Hyperparameters in MultiEMO: The number of layers T in MultiAttn is tuned to be 6, the temperature parameter τ, the sample-weight parameter α and the focusing parameter γ in the SWFC loss are designed to be 0.8, 0.8 and 2 respectively, the combining coefficients μ_1 and μ_2 in the full training loss function L_{Train} are tuned to be 0.4 and 0.3 respectively.

Evaluation Metrics: We use the Weighted-average F1 score (Weighted-F1) for model evaluations.

5 Results and Analysis

5.1 Comparison with Baseline Models

The comparisons between MultiEMO and existing state-of-the-art approaches on IEMOCAP and MELD are shown in Table 1 and Table 2 respectively. Experimental results demonstrate that MultiEMO achieves the new state-of-the-art performances on both datasets and outperforms existing approaches across all emotion categories, with significant improvements in minority and semantically similar classes. Specifically, on IEMOCAP, MultiEMO surpasses MVN by 17.97% Weighted-F1 in the minority class *Happiness* and achieves relative Weighted-F1 improvements of 8.49% and 10.54% in two similar classes *Sadness* and *Frustration* respectively; On MELD, MultiEMO gains a remarkable 153.59% relative improvement in minority emotion *Fear*, and outperforms the previous best baselines in semantically-similar emotion pairs *Anger* and *Disgust* by 11.31% and 68.12%.

5.2 Different Modality Settings

The comparison of MultiEMO with different modality settings on IEMOCAP and MELD is illustrated in Table 3. From Table 3 we can see that the textual modality of utterances plays a major role in ERC, while the complementary cues from audio and visual modalities can bring considerable improvements over the text-based MultiEMO.

5.3 Ablation Study

To study the contributions of different components in MultiEMO to model performances, we conduct ablation studies on both IEMOCAP and MELD, the results are shown in Table 1 and Table 2.

Impact of VisExtNet: To study the effect of VisExtNet, we implement MultiEMO$_{\text{w/o VisExtNet}}$, in which the proposed VisExtNet is replaced by a 3D-CNN. Experimental results show that the performances of MultiEMO$_{\text{w/o VisExtNet}}$ decrease in all emotion categories on both IEMOCAP and MELD, with a more notable decline on MELD, since the complicated multi-party conversations in MELD make it more challenging for a 3D-CNN to accurately capture visual cues. The inferior performances of MultiEMO$_{\text{w/o VisExtNet}}$ on both datasets prove the effectiveness of VisExtNet.

Impact of MultiAttn: To analyze the impact of MultiAttn, we implement MultiEMO$_{\text{w/o MultiAttn}}$, where we replace MultiAttn with feature concatenation to fuse contextualized multimodal features. As shown in Table 1 and Table 2, the performances of MultiEMO$_{\text{w/o MultiAttn}}$ fall sharply in all emotion classes of IEMOCAP and MELD, which proves the importance and superiority of capturing cross-modal correlations and dependencies across textual, audio and visual modalities using MultiAttn.

Impact of SWFC Loss: To study the contribution of SWFC loss, we implement another variant MultiEMO$_{\text{w/o SWFC loss}}$ by removing the SWFC loss part from the training loss function. Experimental results demonstrate that the performances of MultiEMO$_{\text{w/o SWFC loss}}$ drop considerably on both IEMOCAP and MELD, the declines in minority and semantically similar emotion classes are remarkably striking, while the decreases in majority classes are merely marginal. In addition, the degree of decline is more noticeable on MELD since MELD is significantly more class-imbalanced than IEMOCAP. The results of MultiEMO$_{\text{w/o SWFC loss}}$ prove the effectiveness of SWFC loss in mitigating the difficulty of classifying minority and semantically similar emotion categories.

5.4 Case Study

A case study is illustrated in Appendix A.1.

6 Conclusion

In this paper, we propose a novel attention-based correlation-aware multimodal fusion framework named MultiEMO for the task of ERC, in which we design a visual feature extractor VisExtNet to accurately capture emotion-rich visual cues and introduce a multimodal fusion model MultiAttn to effectively model the cross-modal interactions and mapping relationships across multiple modalities. Furthermore, the difficulty of classifying minority and semantically similar emotions is mitigated by our proposed SWFC loss. Extensive experiments on IEMOCAP and MELD demonstrate the effectiveness and superiority of MultiEMO.

Limitations

Although our proposed MultiEMO framework has achieved state-of-the-art performances on both IEMOCAP and MELD, there are some limitations with this work:

- Our proposed visual feature extractor VisExtNet does not distinguish between speakers and irrelevant people in the scene, which can be problematic in some scenarios. For instance, one scene in MELD is the cafeteria, where a lot of background actors sit and drink coffee. The facial expressions of these background people have no impact on the emotion of the speaker since they do not participant in the conversation. However, VisExtNet captures visual features of everyone appeared in the cafeteria with no differentiation, which may lead to a wrong comprehension of the speaker's emotional tendency due to the effects of facial expressions from irrelevant people. We plan to explore effective ways to distinguish between interlocutors and irrelevant people in the scene in our future work.
- The effects of hyperparameters in the SWFC loss (temperature parameter τ, sample-weight parameter α and focusing parameter γ) on model performances have not been fully studied, which will be thoroughly analyzed in our future research.
- Due to the class imbalanced issue with MELD, the SWFC loss requires a large batch size on MELD to ensure that for each training sample there exists at least one positive pair in the batch, which can be computationally expensive. We will investigate effective approaches to tackle this challenge in our future research.
- Even though MultiEMO has achieved remarkable improvements in minority emotion categories, the performances of MultiEMO in minority emotions are still worse than majority classes. How to further improve performances in low-resource emotion classes will be explored in the future.

Ethics Statement

The significant improvements in classifying minority emotion categories brought by our method can make MultiEMO a powerful tool in psychopathological fields such as depression detection, where minority emotions *sadness*, *fear* and *anger* are important early indicators of depression (O'Connor et al., 2002).

Acknowledgements

The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202 and the Shenzhen Science and Technology Program under Grant KQTD20170810150821146.

References

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(4):335–359.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander 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, pages 2470–2481, Online. Association for Computational Linguistics.

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

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

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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 (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. 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 (Volume 1: Long Papers)*, pages 5666–5675, Online. Association for Computational Linguistics.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8002–8009.

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

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

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*.

Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States. 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*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021b. Quantum-inspired neural network for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13270–13278.

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

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hui Ma, Jian Wang, Hongfei Lin, Xuejun Pan, Yijia Zhang, and Zhihao Yang. 2022. A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, 236:107751.

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.

Lynn E O'Connor, Jack W Berry, Joseph Weiss, and Paul Gilbert. 2002. Guilt, fear, submission, and empathy in depression. *Journal of affective disorders*, 71(1-3):19–27.

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 (volume 1: Long papers)*, pages 873–883.

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 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang. 2019. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5281–5288.

Zhiwei Yang, Jing Ma, Hechang Chen, Yunke Zhang, and Yi Chang. 2021. HiTRANS: A hierarchical transformer network for nested named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 124–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5415–5421. International Joint Conferences on Artificial Intelligence Organization.

Haidong Zhang and Yekun Chai. 2021. COIN: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 12–18, Mexico City, Mexico. Association for Computational Linguistics.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. 2021. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Advances in Neural Information Processing Systems*, 34:29848–29860.

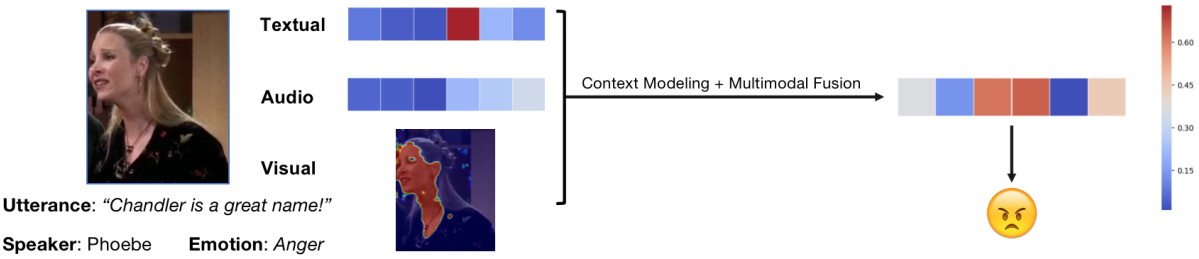


Figure 6: Visualization of the heatmaps of a prone-to-misclassification utterance in MELD.

A Appendix

A.1 Case Study of MultiEMO

Since the one-stage paradigm (Section 3.3.1) simultaneously performs unimodal textual feature extraction and textual context modeling, to better illustrate the role of context modeling to emotion classification, in the section of case study, the textual modality of the selected utterance is processed using a two-stage paradigm (Yang et al., 2021): unimodal feature extraction with a pre-trained RoBERTa and context modeling with another transformer [1], such that the impact of context modeling on the textual modality can be analyzed in conjunction with audio and visual modalities.

Figure 6 depicts a visualization of a prone-to-misclassification utterance in MELD, in which the textual modality "*Chandler is a great name!*" appears to be positive while the true connotation of the utterance actually implies anger. The heatmaps of the utterance's textual, audio and visual modalities on the left are obtained after unimodal feature extractions, from which we can see that: (1) Textual modality: the word "great" plays a major role in the text, revealing a strong positive emotion; (2) Audio modality: the higher intensity in the latter part of the audio indicates a flat-to-sharp tone; (3) Visual modality: the frown in the speaker's face implies a negative emotion. The asynchronization of emotional tendencies from different modalities makes it challenging to identify the actual emotion of this utterance. However, by modeling contextual information and capturing complex cross-modal correlations across contextualized textual, audio and visual modalities, MultiEMO learns a highly representative feature for this utterance, as shown in the heatmap on the right of Figure 6. The learned multimodal-fused feature can be easily classified to the correct emotion class since it preserves useful emotional cues while discarding irrelevant information through selectively focusing on highly-correlated information across contextualized textual, audio and visual modalities.

[1] As mentioned in Section 3.3.1, the performance of MultiEMO with a two-stage paradigm is merely marginally outperformed by the one-stage paradigm, both approaches can learn good contextualized textual representations.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The section of Limitations is after the Conclusion section.

☒ A2. Did you discuss any potential risks of your work?
We do not discuss potential risks of our work because of the page limit.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The sections of Abstract and section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ 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.
No response.

C ☑ Did you run computational experiments?

Section 4 and Section 5.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We do not report the information mentioned in the question because of the page limit.

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?
We do not report the information mentioned in the question because of the page limit.

☒ 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.)?
We do not report the information mentioned in the question because of the page limit.

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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.