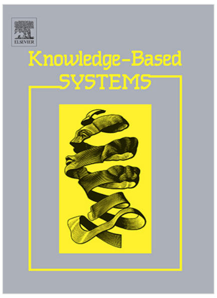
Journal Pre-proof



Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation

ShiHao Zou, Xianying Huang, XuDong Shen, Hankai Liu

PII: S0950-7051(22)01071-1
DOI: https://doi.org/10.1016/j.knosys.2022.109978
Reference: KNOSYS 109978

To appear in: *Knowledge-Based Systems*

Received date : 16 July 2022
Revised date : 28 September 2022
Accepted date : 30 September 2022

Please cite this article as: S. Zou, X. Huang, X. Shen et al., Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation, *Knowledge-Based Systems* (2022), doi: https://doi.org/10.1016/j.knosys.2022.109978.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Elsevier B.V. All rights reserved.

Journal Pre-proof

Revised Manuscript (Clean Version) Click here to view linked References

Improving Multimodal Fusion with Main Modal Transformer for Emotion Recognition in Conversation

ShiHao Zou, Xianying Huang*, XuDong Shen and Hankai Liu

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

ARTICLE INFO

Keywords:
Emotion recognition in conversation
Main modal
Transformer
Multihead attention
Emotional cues

ABSTRACT

Emotion recognition in conversation (ERC) is essential for developing empathic conversation systems. In conversation, emotions can exist in multiple modalities, i.e., audio, text, and visual. Due to the inherent characteristics of each modality, it is not easy for the model to use all modalities effectively when fusing modal information. However, existing approaches consider the same representation ability of each modality, resulting in unsatisfactory fusion across modalities. Therefore, we consider different modalities with different representation abilities, propose the concept of the main modal, i.e., the modal with stronger representation ability after feature extraction, and then propose the method of Main Modal Transformer (MMTr) to improve the effect of multimodal fusion. The method preserves the integrity of the main modal features and enhances the representation of weak modalities by using multihead attention to learn the information interactions between modalities. In addition, we designed a new emotional cue extractor that extracts emotional cues from two levels(the speaker's self-context and the contextual context in conversation)to enrich the conversation information obtained by each modal. Extensive experiments on two benchmark datasets validate the effectiveness and superiority of our model.

1. Introduction

The main goal of emotion recognition in conversation (ERC) is to correctly identify the emotions expressed by each speaker's utterance during a conversation. Recently, there has been an increasing number of works on building intelligent responses in dialog systems such as open-domain [1], task-oriented [2], and fusion of open-domain and task-oriented [3]. ERC has greatly increased its importance in constructing a dialog system that can understand the users' emotions and intentions and conduct effective dialog interactions as a relevant task for dialog systems. It has made an essential contribution to the better development of engaging, interactive, and empathetic dialog systems [4], which has greatly advanced the development of human-computer interaction. Especially in the current situation of novel coronavirus pneumonia, the above research is more relevant.

Studies [5] have shown that humans prefer to feel emotions better through multiple modalities, such as visual and audio. Figure 1 shows a sample of emotional expressions in three different modalities. The input to the multimodal ERC is different modal information for each utterance, and the model uses this information to make relevant emotion predictions for the utterance. Since the conversation contains rich emotional cues, such as speaker information and conversation context information, how to effectively utilize the emotional cues among the modalities is an urgent problem.

Early research on ERC fused multimodal information by manipulating feature tensors. Zhu et al. [6] performed multimodal fusion by decomposing the tensor and weights in parallel and using modal-specific low-order factors. With the development of graph convolutional networks (GCNs), there is a significant improvement in the performance of GCN-based models on multimodal ERC compared to previous models. For example, Hu et al. [7] proposed a multimodal fused graph convolutional network (MMGCN) to model multimodal information and simultaneously capture long-distance contextual information effectively. However, the existing multimodal ERC research works consider the representation ability of different modalities as the same and thus enhance the modal features through the information interaction between the modalities. Since the difference in the representation ability of the modalities is not a consideration, which leads to the introduction of many noises during the information interaction and weakens the representation ability of the modalities, the effect of multimodal fusion in the above method is not satisfactory.

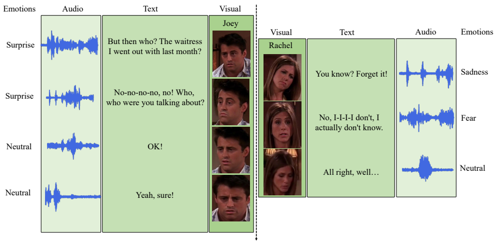


Figure 1: Multimodal conversation example in MELD.

To address the above issues, we define the modal with stronger representational ability as the main modal and use different learning methods for modalities with different representational abilities. Based on previous studies [8, 9, 10]

*Corresponding author.
E-mail addresses: z_sh9904@163.com (S. Zou), wldsj_cqut@163.com (X. Huang), shenxudong0901@163.com (X. Shen), lhk_cqut@163.com (H. Liu)
ORCID(s): 0000-0002-6648-3995 (S. Zou)

Journal Pre-proof

Journal Pre-proof

on psychology, language is the most successful communication tool, enabling the propagation and accumulation of human civilization. Language plays a constitutive role in emotion perception because words ground the otherwise highly variable instances of an emotion category and are brought to bear to make meaning of facial expression movements in a given context [11]. Hence, we chose to use the text representing language as the main modal. In this paper, we propose the main modal transformer (MMTr) to improve multimodal fusion methods for a better understanding of ERC tasks. MMTr is a transformer-based model but does not contain an encoder-decoder part. It only uses multiple single- and multihead attention mechanisms that cross-modally learn features of the source modality to enhance the representation of the target modality. First, we use two levels of bidirectional long short-term memory(Bi-LSTM) for the text modality to extract emotional cues at the speaker's self-context level and contextual context-level, i.e., the cues in the utterance by the same speaker and the emotional cues between the whole conversation context. Then, the emotional cues are fused to obtain the purified text feature representation. For visual and audio modalities, contextual context-level Bi-LSTM extracts the corresponding emotional cues. It effectively alleviates the problem of extracting and fusing emotional cues. Second, we select the text modality as the main modality and the audio and visual modalities as the target modalities. The target modality continuously learns the features of other modalities to obtain enhanced feature representations. MMTr enables the main modality to maintain its complete representational capability, while the relatively weaker modality gradually increases its representative ability by learning from other modalities, which effectively alleviates the problem of representational decay occurring during modal fusion. Finally, the learned cross-modal features are fused and used to obtain a feature representation for emotion classification.

We conducted a series of experiments on two public benchmark multimodal datasets, and the results consistently show that MMTr significantly outperforms various baseline methods. The main contributions of this work are as follows:

- We propose that the representational abilities of each modal in a multimodal task are not the same, and modalities with different representational abilities should be learned differently,i.e.,from multimodal to select the main modal, and then different learning is performed.
- The MMTr model is proposed to facilitate contextual understanding of multimodal ERC by preserving the integrity of the main modal features while enhancing the weak modal representations to effectively perform intermodal information fusion.
- Extensive experimental results on two public benchmark multimodal datasets demonstrate that our proposed MMTr outperforms the state-of-the-art baseline models.

2. Related Work

ERC differs from traditional emotion recognition tasks. Instead of considering emotions as static states, ERC is a state that constantly changes with the conversation, where context plays a crucial role. Previous work on ERC has mainly used text [12], and in the last few years, datasets with visual, text, and audio cues have been made publicly available [13, 14]. On these datasets, multiple deep learning methods are applied to identify emotions. We classify them as either only using text or using multimodal data.

2.1. Text-based Methods

As an important research area in natural language processing, ERC has received extensive attention in recent years. Ghosal et al. [15] presented a dialog graph convolutional network (DialogueGCN) that adopts a GCN to capture the self- and interspeaker dependencies among utterances, which effectively solves the context propagation problem of DialogueRNN [16]. Ishiwatari et al. [17] proposed relational graph attention network(R-GAT) with relational position encoding not only captures the dependencies between speakers, but also provides sequential information about the structure of the relational graph. Ma et al. [18] designed a hierarchical attention network with a residual gated recurrent unit (HAN-ReGRU) framework to capture the long range contextual information in an utterance and a conversation. Li et al. [19] proposed a bidirectional emotional recurrent unit (BiERU) framework that uses a generalized neural tensor block to perform context compositionality and employs an emotion feature extractor to yield emotional features. Ma et al. [20] adopted a multiview network(MVN) to explore the emotion representation of a query from two different views. Zhu et al. [21] utilized the Encoder-Decoder architecture, which combines the representation of topic information with common-sense information in ERC. Yang et al. [22] introduced curriculum learning into the field of ERC for the first time by setting two levels of curriculum to divide the data. Based on some of the models mentioned above, the performance of the model was greatly improved.

The above approaches suggest that contextual and speaker information in conversation is beneficial for emotion recognition. However, the existing approach may be affected when facing a lack of future information in the real world or sudden changes in emotion during a conversation.

2.2. Multimodal-based Methods

Multimodal ERC collects and processes data from multiple sources (e.g., audio, visual, text, and other information) to understand various human emotions from multiple aspects [23]. Hazarika et al. [24] proposed a conversational memory network(CMN) that employs separate memory networks to store the contextual information for both interlocutors. To improve CMN, Hazarika et al. [25] presented an interactive conversational memory network(ICON) to hierarchically model the self- and inter-speaker emotional influences into global memories. Tsai et al. [26] introduced

Journal Pre-proof

Journal Pre-proof

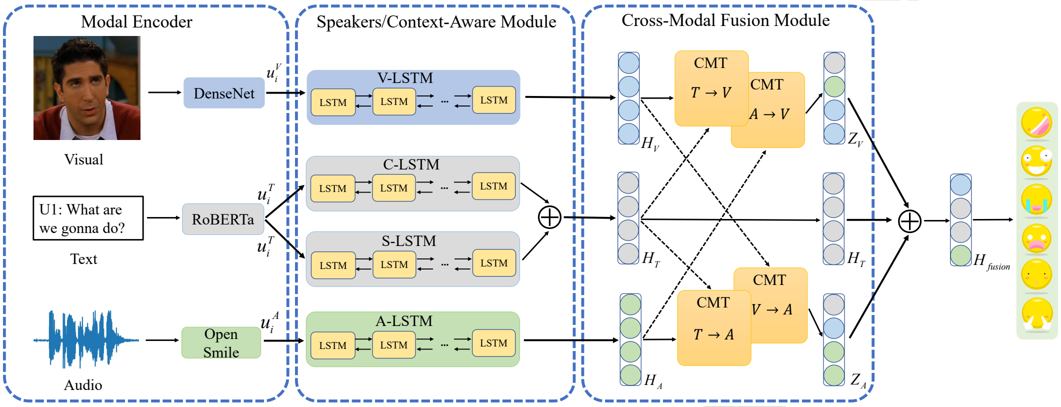


Figure 2: Overall architecture of the proposed MMTr.

the multimodal transformer(MulT) to use the basic module of the transformer-fusion method with multihead attention mechanism, which achieves cross-modal information fusion by using different modalities as Query, Key, and Value in attention, respectively. Liu et al. [27] designed a representation learning method for multimodal data using contrast loss that learns the complementary synergy between modality effects. Han et al. [28] presented a multimodal infomax (MMIM) to maintain task-relevant information by maximizing mutual information in unimodal input pairs. Tu et al. [29] proposed a multitask graph neural network (MGNN) to implement a cooperative mechanism in discrete and dimensional models. This mechanism not only enables the emotion recognition model to accurately locate simple and discrete emotional anchors in the entire continuous emotion space (exploration) but also encourages it to effectively search for complex and subtle emotional states near the emotional anchors(exploitation). Hu et al. [30] proposed a new multimodal dynamic fusion network(MM-DFN) to capture dynamic changes in contextual information in different semantic spaces.

The above studies show that multimodal features have better performance and robustness than unimodal features, which is more evident in the emotion recognition task.

3. Problem Formulation

In ERC, a conversation is defined as a sequence of utterances $\{u_1, u_2, \dots, u_N\}$, where N is the number of utterances in the conversation. Each utterance u_i consists of n_i tokens, i.e $\{w_{i1}, w_{i2}, \dots, w_{in_i}\}$. Each conversation has M speakers $P = \{p_1, p_2, \dots, p_M\}, (M \geq 2)$, and each utterance u_i is spoken by a speaker $p_{\phi(u_i)}$, where the utterance index ϕ is mapped to the corresponding speaker. In addition, we define U_λ to represent the set of utterance spoken by the party p_λ. $U_\lambda = \{u_i | u_i \in U$ and u_i spoken by $p_\lambda, \forall i \in [1, N]\}$, $\lambda \in [1, M]$. The discrete values $y_i \in$ S are used to represent the emotion labels of u_i, where S is the set of emotion labels. The objective of this task is to predict the emotion label y_i for a given query utterance u_i based on dialog context $\{u_1, u_2, \dots, u_N\}$ and the corresponding information. Each utterance contains data sources from three features corresponding to three modalities Audio (A), Visual (V), and Text (T), denoted as follows:

$$u_i = [u_i^A, u_i^V, u_i^T], \tag{1}$$

where $u_i^A \in \mathbb{R}^{d_A}$, $u_i^V \in \mathbb{R}^{d_V}$,and $u_i^T \in \mathbb{R}^{d_T}$ represent the Audio, Visual, and Text modalities,respectively. For the rest of the paper,$d_{(\cdot)}$ represents the feature dimension.

4. Proposed MMTr

Our MMTr modeling is as follows: we obtain three modal features of the conversation data: text, audio, and visual. We obtain the three modal features of the emotional cues and use the cross-modal fusion(CMF) module to fuse the modal features through different learning, and finally obtain the results of emotional recognition. Figure 2 shows the overall architecture of the MMTr. Note that the text modality is the main modality in our model, so we use the Speakers/Context-Aware (S/C-Aware) module to obtain two levels of emotional cues and use the contextual context-level emotional cue extractor for audio and visual modalities to obtain more conversational information.

4.1. Unimodal Feature Extraction

4.1.1. Text feature extraction

To obtain better utterance representation and achieve our goal of obtaining modal features with strong representational ability, we use the large general pretrained language model RoBERTa-Large [31] for text vector encoding extraction. The architecture of RoBERTa-Large is the same as BERT-Large [32]. Based on BERT, it has been optimized by using larger batches of more data and training the model

Journal Pre-proof

Journal Pre-proof

for a longer time. However, unlike other downstream tasks, we use the transformer structure to encode the utterances without classifying or decoding them. More specifically, for each utterance in the text modal, we precede its token with a special token $[CLS]$ to make it of the form of $\{[CLS], w_{i1}, w_{i2}, \ldots, w_{in_i}\}$. Then we use the pooled embedding result of the last layer of $[CLS]$ as the feature representation of u_i^T, and finally, we obtain a sentence vector with 1024 dimensions for each utterance.

4.1.2. Audio feature extraction

According to the configuration of ICON [25], we used OpenSmile [33] for audio feature extraction. With the IS13 comparison profile, which extracted a total of 6373 features for each utterance video, we reduced the dimensionality to 1582 for the IEMOCAP and 300 for the MELD dataset by using a fully connected layer.

4.1.3. Visual feature extraction

The visual facial features were extracted by pretraining on the Facial Expression Recognition Plus (FER+) [34] corpus using DenseNet [35]. This captures changes in the expression of the speakers, which is very important information for ERC. Finally, a 342-dimensional visual feature representation was obtained.

4.2. Speaker/Context-Aware Module

Existing methods mainly use GRU, LSTM, and attention to extract emotional cues. In this part, two Bi-LSTMs are used to capture emotional cues at the contextual context-level and the speaker's self-context level.

4.2.1. C-LSTM module

To learn the contextual representation at the contextual context-level, we capture the sequential dependencies between neighboring utterances in the conversation context by C-LSTM. Taking the textual features of each utterance $\{u_i^T\}_{i=1}^N \in \mathbb{R}^{d_T}$ as input, the contextual context-level cues representation $t_i^c \in \mathbb{R}^{2d_T}$ can be computed as follows:

$$t_i^c, h_i^c = \overleftrightarrow{LSTM}^C(u_i^T, h_{i-1}^c), \tag{2}$$

where $h_i^c \in \mathbb{R}^{d_T}$ is the i-th hidden layer state of the LSTM in the context layer.

4.2.2. S-LSTM module

To learn contextual representations at the speaker's self-context level, we use a S-LSTM to capture correlations between neighboring utterances of the same speaker. Given the textual features $\{u_i^T\}_{i=1}^N$ of each utterance, the speaker's self-context level cues representation $c_i^s \in \mathbb{R}^{2d_T}$ can be computed as:

$$c_i^s, h_{\lambda,j}^s = \overleftrightarrow{LSTM}^S(u_i^T, h_{\lambda,j-1}^s), j \in [1, |\mathrm{U}_\lambda|], \tag{3}$$

where $\lambda = \phi(u_i)$, U_λ refers to the set of all utterances of the speaker p_λ. $h_{\lambda,j}^s \in \mathbb{R}^{d_T}$ is the hidden layer state of the j-th speaker's self-context level LSTM of speaker p_λ .

Finally, the two levels of emotional cues obtained from the above calculation are fused to obtain an information-enhanced textual modal representation H_i^T of the current utterance, which is calculated as follows:

$$H_i^T = t_i^c + c_i^s, \tag{4}$$

4.2.3. A/V-LSTM module

For audio and visual modal, we use A/V-LSTM for contextual context-level cues extraction, which is computed as follows:

$$\begin{aligned} H_i^A, h_i^a &= \overleftrightarrow{LSTM}^A(u_i^A, h_{i-1}^a), \\ H_i^V, h_i^v &= \overleftrightarrow{LSTM}^V(u_i^V, h_{i-1}^v), \end{aligned} \tag{5}$$

where h_i^a , h_i^v are divided into the i-th hidden layer states of the audio and visual modal LSTM, and $H_i^T \in \mathbb{R}^{2d_T}$, $H_i^A \in \mathbb{R}^{2d_A}$, $H_i^V \in \mathbb{R}^{2d_V}$ are the contextual feature representations of text, audio, and visual, respectively.

Finally, the individual modal features of the N utterances in the conversation are aggregated separately, and the features of each modal are aligned through a linear layer to d_T:

$$\begin{aligned} H_T &= \mathrm{Linear}(\|_{i=1}^N \ H_i^T), \\ H_A &= \mathrm{Linear}(\|_{i=1}^N \ H_i^A), \\ H_V &= \mathrm{Linear}(\|_{i=1}^N \ H_i^V), \end{aligned} \tag{6}$$

where $\|$ denotes the concatenation operation, $H_T, H_A, H_V \in \mathbb{R}^{L_T \times d_T}$,and $L_{(\cdot)}$ is the sequence length.

4.3. Cross-modal Fusion Module

4.3.1. Cross-modal attention

Cross-modal attention is used to achieve cross-modal information interaction in the context of utterances, by which the module potentially fuses the information flow from the source modal to the target modal, e.g., $Text \rightarrow Visual$, and enhances the representation of the visual modal by learning the feature representation of the text modal.

First, we define Query as $\mathrm{Q}_V = H_V W_{\mathrm{Q_V}}$, Keys as $\mathrm{K}_T = H_T W_{\mathrm{K_T}}$, and Values as $\mathrm{V}_T = H_T W_{\mathrm{V_T}}$, where $W_{(\cdot)}$ is the trainable weight matrix. The information flow fusion from text modal to visual modal is represented on the cross-modal attention as:

$$\begin{aligned} \mathrm{Y}_{VT} &= \mathrm{CA}_{T\rightarrow V}(H_V, H_T) \\ &= softmax(\frac{\mathrm{Q}_V \mathrm{K}_T^T}{\sqrt{d_T}})\mathrm{V}_T, \end{aligned} \tag{7}$$

where $\mathrm{Y}_{VT} \in \mathbb{R}^{L_T \times d_T}$. CA is the cross-modal attention module, which is used to compute the score matrix between the two modalities. We learn adaptively from the low-level features, which facilitates our model to retain the low-level information of each source modal.

Journal Pre-proof

4.3.2. *Cross-modal transformer*

Based on the above cross-modal attention, we designed a cross-modal transformer (CMT) with the structure shown in Figure 3. The learning of this module makes the modality with weaker representation ability enhance the representation ability of its modality by learning the features of other modalities. In the following example, we take the text information transfer to the visual as an example, denoted as $T \rightarrow V$. Each CMT is composed of a cross-modal attention block at layer D. A CMT is computed on the previous D layer as follows:

$$\begin{aligned} Z_{T\rightarrow V}^{(0)} &= H_V^{(0)}, \\ \tilde{Z}_{T\rightarrow V}^{(i)} &= \mathrm{CA}_{T\rightarrow V}^{(i),mul}(\mathrm{Norm}(Z_{T\rightarrow V}^{(i-1)}), \mathrm{Norm}(H_T^{(0)})) \\ &+ \mathrm{Norm}(Z_{T\rightarrow V}^{(i-1)}), \qquad (8) \\ Z_{T\rightarrow V}^{(i)} &= \mathrm{f}_{\theta_{\mathrm{T}\rightarrow\mathrm{V}}^{(\mathrm{i})}}(\mathrm{Norm}(\tilde{Z}_{T\rightarrow V}^{(i)})) + \mathrm{Norm}(\tilde{Z}_{T\rightarrow V}^{(i)}), \qquad (9) \end{aligned}$$

where f_θ is a position forward sublayer parameterized by θ , and $\mathrm{CA}_{(\cdot)}^{mul}$ denotes the multihead cross-modal attention block.

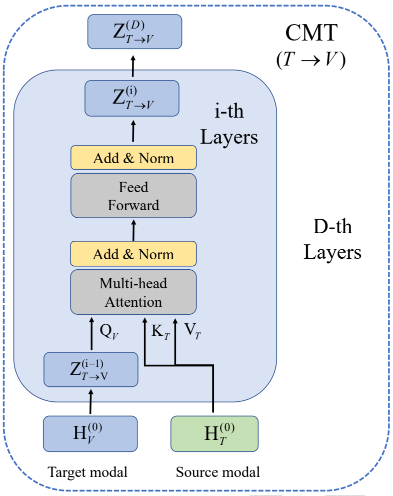


Figure 3: The architecture of CMT.

In this process, each modal continuously updates its representational abilities through low-level external information from the multihead cross-modal attention block. At each level of the cross-modal attention block, the source modality's low-level information is transformed into a different set of Key/Value pairs that interact with the target modality.

4.3.3. *Self-attention*

We aggregate the output of MMTr sharing the same target modal, i.e., $X_{\{V,A\}} \in \mathbb{R}^{L_T \times d_T}$ is calculated as in Equation(10). Then, the sequence information is collected by self-attention to obtain the feature representation of a modal of the current utterance after fusion.

$$X_V = Z_{T\rightarrow V}^D \oplus Z_{A\rightarrow V}^D, \qquad (10)$$

$$Z_V = \mathrm{Attention}(X_V), \qquad (11)$$

4.4. Modal Fusion

We connect the fused target modality $Z_{(\cdot)}$ with the main modal H_T and generate feature representations X_{fusion} from different modalities for each utterance. Finally, we collect the comprehensive sequence information of X_{fusion} through self-attention to obtain the final feature representation H_{fusion} of the current utterance.

$$X_{fusion} = H_T \oplus Z_V \oplus Z_A, \qquad (12)$$

$$H_{fusion} = \mathrm{Attention}(X_{fusion}), \qquad (13)$$

Then, H_{fusion} is fed into an MLP with a fully connected layer for predicting the emotion label $\hat{y}_i$.

$$P_i = \mathrm{RELU}(W_L H_{fusion} + b_L), \qquad (14)$$

$$\hat{y}_i = \underset{k}{argmax}(P_i[k]) \qquad (15)$$

4.5. Loss Function

We use the standard cross-entropy and L2 regularization as the loss function in the training process.

$$\mathcal{L} = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} log P_{i,j}[y_{i,j}] + \lambda \|\theta\|_2 \qquad (16)$$

where N is the number of conversations, $c(i)$ is the number of utterances in conversation i, $P_{i,j}$ is the probability distribution of the predicted emotion labels of utterance j in conversation i, $y_{i,j}$ is the predicted category labels of utterance j in conversation i, λ is the L2 regularization weight, and θ is the set of all trainable parameters. We use the Adam [36] optimizer with stochastic gradient descent to train our network model.

5. EXPERIMENTS

In this section, we present the experimental setting consisting of datasets, evaluation metrics,baselines,and implementation details.

5.1. Datasets and Evaluations

We evaluated the effectiveness of our model on two benchmark datasets: IEMOCAP [14] and MELD [13]. Both datasets are multimodal ERC datasets containing text, audio, and visual. We divided the datasets according to MMGCN [7] . Table 1 shows the data distribution of the two datasets.

- **IEMOCAP**:The multimodal ERC dataset. Each conversation in IEMOCAP is from two actors' performances based on the script. There are 7433 utterances and 151 conversations in IEMOCAP. Each utterance in the conversation is labeled with six categories of emotions: *happy, sad, neutral, angry, excited*, and *frustrated*.

Journal Pre-proof

Table 1
Data distribution of IEMOCAP and MELD.

Dataset	♯Conversation		♯Utterance		♯Classes
	Train+Val	Test	Train+Val	Test	
IEMOCAP	120	31	5810	1623	6
MELD	1153	280	11098	2610	7

- **MELD**:The data were obtained from the TV show *Friends* and included a total of 13708 utterances and 1433 conversations. Unlike the IEMOCAP dyadic dataset, MELD has three or more speakers in a conversation, and each utterance in the conversation is labeled with seven categories of emotions: *neutral, surprise, fear, sadness, joy, disgust*, and *anger*.

Evaluation metrics: We use the F1-score to evaluate the performance for each emotion class and use the weighted average of accuracy and F1-score to evaluate the overall performance on the two datasets.

5.2. Baseline Model

- **BC-LSTM** [37]: It encodes contextual semantic information through a Bi-LSTM network, but does not consider the speaker's information.
- **ICON** [25]: Two GRUs are used to model the speaker's information, additional global GRUs are used to track changes in emotional states throughout the conversation, and a multilayer memory network is used to model global emotional states. However, the ICON still cannot be adapted to multiple speakers scenarios.
- **DialogueRNN** [16]: It models the speaker and sequential information in a conversation through three different GRUs (global GRU, speaker GRU, and emotion GRU), but DialogueRNN does not improve much in the multimodal domain.
- **DialogueGCN** [15]: It applies GCN to ERC, and the generated features can integrate rich information. RGCN and GCN are both nonspectral domain GCN models for encoding graphs.
- **DialogueXL** [38]: DialogueXL uses the XLNet model for ERC to obtain global contextual information.
- **DialogueCRN** [39]: DialogueCRN introduces a cognitive phase that extracts and integrates emotional cues from the context retrieved during the perception phase.
- **MMGCN** [7]: MMGCN uses GCN networks to obtain contextual information, which not only effectively compensates for the drawback of not being able to exploit multimodal dependencies in DialogueGCN but also effectively uses the speaker's information for ERC.
- **MM-DFN** [30]: MM-DFN fuses multimodal contextual information by designing a new graph-based dynamic fusion module to fully understand multimodal conversational contexts to recognize emotions in utterances.

5.3. Implementation Details

We implemented our proposed MMTr model on the PyTorch framework. The hyperparameters are set as follows: the number of multihead attention heads in MMTr is 5 (i.e., $m = 5$), where the number of layers in the multihead attention module is 5 (i.e., $l = 5$). Dropout was 0.2 in both the IEMOCAP and MELD. The learning rate is 0.0001 in IEMOCAP and 0.0003 in MELD. The L2 regularization is set to 3e-05. The batch_size on both IEMOCAP and MELD is set to 16 . Each training and testing process runs on a single RTX 3090 GPU, with each training process containing 25 epochs in IEMOCAP and up to 2.5 seconds per epoch and 20 epochs in MELD and up to 15 seconds per epoch. The reports of our implemented models are based on the average scores of 5 random runs on the test set.

6. Results and Analysis

We discuss the experimental results of our proposed and baseline models and conduct an ablation study to investigate the contributions of the emotional cue extractor and the main modal. Then, we verify the effects of text embedding, various modalities, and parameters on the model through different experiments. Finally, we perform error analysis and a case study.

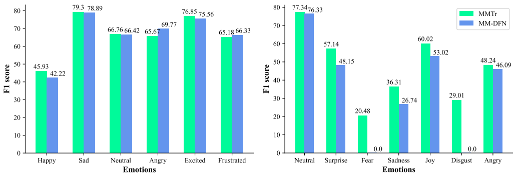
6.1. Comparison with Other Baseline Models

Table 2 shows our proposed MMTr and baseline experimental results on the IEMOCAP and MELD datasets. The baseline results followed by "∗" are rerun using the open-source code. For a fair comparison, using our processed data, we performed experiments on all baseline models that we could reconstruct to later compare the effects of text embedding, shown in the table as "+RoBERTa". "-" means that these results are unavailable in the original paper. Other baselines with results were copied from [30].

Table 2 and Figure 4 report the experimental results on the IEMOCAP and MELD dataset. We find that: (1) The proposed MMTr outperforms all the baseline models in terms of the weighted accuracy and F1-score, demonstrating the effectiveness of our model on multimodal ERC. (2) MMTr outperforms MM-DFN regarding the weighted accuracy and F1-score. This indicates that MMTr has better results for extracting the speaker's information in conversations than the state-of-the-art baseline model that uses the speaker's information. (3) In the comparison between MMTr and MM-DFN for individual emotion categories, as shown in Figure 4, we obtained the best performance for all categories in the MELD dataset and achieved the best results for most of the emotion categories in the IEMOCAP dataset. In particular, among the emotion categories in MELD, many of them have smaller sample numbers, such as surprise, joy, fear,

Journal Pre-proof

disgust, which are more difficult to classify as emotion categories and achieve much better results than MM-DFN. Note that MM-DFN also reports the F1-score per class, except for two classes (i.e.,Fear and Disgust) on MELD, whose results are not statistically significant due to the smaller number of training samples,so they are combined into other similar emotion category samples. These results verify that the proposed model recognizes the most emotional classes, including the minority classes.



(a) IEMOCAP (b) MELD

Figure 4: Model-predicted emotional category scores on IEMOCAP and MELD between MMTr and MM-DFN.

6.2. Ablation Study

To investigate the contributions of the emotional cues extractor and the main modal proposed in MMTr, we conduct an ablation study on the two datasets. We consider the following settings:

- **Only T**:We use only the text modal for multimodal ERC.
- **Only A**:We use only the audio modal for multimodal ERC.
- **Only V**:We use only the visual modal for multimodal ERC.
- **A as the main modal**:We use the audio modal as the main modal for multimodal ERC.
- **V as the main modal**:We use the visual modal as the main modal for multimodal ERC.
- **T+A+V**:We treat the representational abilities of the three modalities as the same and use them for multimodal information fusion.
- **w/o S/C-Aware**:We remove the emotion cues extractor used.

Table 2
Experimental results on IEMOCAP and MELD datasets. Results of baselines followed by "∗" are rerun using open-source code ,and other baselines are copied from [30]. Best results are in boldface, and "-" means that these results are unavailable from original paper.

Models	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
BC-LSTM*	60.94	60.01	59.27	56.97
ICON*	64.00	63.50	-	-
DialogueRNN*	64.26	63.14	59.81	57.59
+RoBERTa	66.42	66.39	63.41	62.86
DialogueGCN*	63.22	62.89	60.31	56.36
+RoBERTa	66.05	64.91	62.99	62.76
DialogueXL*	-	65.94	-	62.41
DialogueCRN*	67.16	67.21	61.11	58.67
+RoBERTa	68.15	68.35	63.87	63.69
MMGCN*	66.06	65.65	61.26	57.97
+RoBERTa	69.13	69.01	64.18	63.54
MM-DFN	68.21	68.18	62.49	59.46
Ours(MMTr)	70.87	69.53	62.50	60.58
+RoBERTa	**72.27**	**71.91**	**64.64**	**64.41**

Table 3 and Figure 5 show the ablation results. From the results, we conclude that: (1) The performance of multimodal data input is better than that of single-modal data input, and the performance of text modal is far better than that of the other two modalities. (2) Comparing only T and T+V+A, we find that the results are better when adding two other modal information to the text features than when using only a single text modality. This is because audio and visual can assist the text to some extent, especially in utterance where the emotional expression of the text is not obvious. However, we found that the performance on MELD is not apparent, found through exploration: the number of utterances per conversation in MELD is higher, and the utterances are shorter, the extraction of audio and visual features corresponding to the utterances is poor, so it leads to no significant improvement after adding the information of these two modalities. (3) By selecting audio and visual as the main modal, we verify the correctness of our text selection, which has stronger representational power as the main modal. (4) The S/C-Aware emotion cue extractor is useful because removing it leads to a performance decrease on the two datasets. (5) Comparing the IEMOCAP and MELD datasets, we find that MELD is more challenging to extract emotional cues in S/C-Aware than IEMOCAP. This is probably because IEMOCAP and MELD are dyadic and multiparty conversation datasets; hence, utilizing the speaker information is more critical for MELD. However, our emotion cue extractor for multiperson scenarios is not yet able to accurately extract the emotion cues of each speaker.

6.3. Effect of Text Embedding

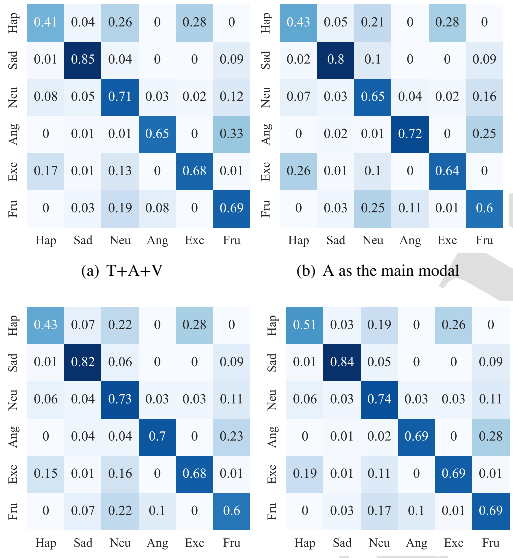
Table 2 reports the experimental results. We observe that using RoBERTa embeddings has better performance than using TextCNN embeddings on the two datasets, both in the baseline model and MMTr. This indicates that high-quality deep contextualized word representations can further improve the effectiveness of the model. The benefit also coincides with our expectation of obtaining the best main modal representation capability after selecting the main modalities. Moreover, compared with TextCNN embeddings, MMTr

Journal Pre-proof

Journal Pre-proof

Table 3
Ablation study on two datasets. T for text modal, A for audio modal, and V for visual modal.

	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
Only T	67.42	67.28	63.40	63.23
Only A	45.78	46.42	43.22	38.85
Only V	38.08	38.25	36.25	34.19
A as the main modal	70.00	69.50	63.28	63.40
V as the main modal	69.64	69.28	63.28	63.41
T+V+A	70.53	70.48	63.39	63.40
w/o S/C-Aware	62.60	62.41	62.42	62.20
Ours	**72.27**	**71.91**	**64.64**	**64.41**



	Hap	Sad	Neu	Ang	Exc	Fru
Hap	0.41	0.04	0.26	0	0.28	0
Sad	0.01	0.85	0.04	0	0	0.09
Neu	0.08	0.05	0.71	0.03	0.02	0.12
Ang	0	0.01	0.01	0.65	0	0.33
Exc	0.17	0.01	0.13	0	0.68	0.01
Fru	0	0.03	0.19	0.08	0	0.69

(a) T+A+V

	Hap	Sad	Neu	Ang	Exc	Fru
Hap	0.43	0.05	0.21	0	0.28	0
Sad	0.02	0.8	0.1	0	0	0.09
Neu	0.07	0.03	0.65	0.04	0.02	0.16
Ang	0	0.02	0.01	0.72	0	0.25
Exc	0.26	0.01	0.1	0	0.64	0
Fru	0	0.03	0.25	0.11	0.01	0.6

(b) A as the main modal

	Hap	Sad	Neu	Ang	Exc	Fru
Hap	0.43	0.07	0.22	0	0.28	0
Sad	0.01	0.82	0.06	0	0	0.09
Neu	0.06	0.04	0.73	0.03	0.03	0.11
Ang	0	0.04	0.04	0.7	0	0.23
Exc	0.15	0.01	0.16	0	0.68	0.01
Fru	0	0.07	0.22	0.1	0	0.6

(c) V as the main modal

	Hap	Sad	Neu	Ang	Exc	Fru
Hap	0.51	0.03	0.19	0	0.26	0
Sad	0.01	0.84	0.05	0	0	0.09
Neu	0.06	0.03	0.74	0.03	0.03	0.11
Ang	0	0.01	0.02	0.69	0	0.28
Exc	0.19	0.01	0.11	0	0.69	0.01
Fru	0	0.03	0.17	0.1	0.01	0.69

(d) Ours

Figure 5: Confusion matrix for IEMOCAP modal combinations.

using RoBERTa embeddings has more significant improvement on MELD than on IEMOCAP(i.e.,2.14% vs. 1.4% and 3.83% vs. 2.38% in terms of the weighted accuracy and F1-score, respectively). The reason, as analyzed in (2) of the ablation experiment, is that since the extracted audio and visual features are not very good, it is necessary to enhance the modal features by the main modal features.Therefore the performance gain obtained by selecting RoBERTa as the text embedding is desirable and necessary.

6.4. Various Modality and Analysis

Table 3 shows the performance of our model on the MELD and IEMOCAP datasets under different modal combinations. It is easy to find that: (1) For audio features, the frequency and amplitude of the voice can reflect the intensity of the speaker's emotion but not the intensity of a specific emotion. For example, the voice when happy and angry may both have a distinct upward intonation compared to other emotions. Therefore, it is difficult to correctly distinguish the current speaker's emotion through audio data alone when certain emotions have similar frequencies and amplitudes. (2) For visual features, it is easy to judge the speaker's expression by facial features, but because the speaker intentionally hides his or her facial expression, it is difficult for visual features to make correct emotional judgments. (3) We verify the influence of the fusion method with different modalities as the main modal on the model, and the three modalities are treated on the model equally. Both achieve good results but still fall short of MMTr. The results indicate that taking text as the main modal has a significant impact. The fusion methods that do not consider the modal representation ability introduce much noise and weaken the modal representation ability.

6.5. Parameter Sensitivity

In the proposed model, we use different numbers of heads in multihead attention and layers of cross-modal multihead attention and observe the corresponding w-F1 scores. This analysis is demonstrated in Figure 6, where one can observe that the proposed model with $m = 5$ heads in the cross-modal attention module and using $l = 5$ layers of cross-modal attention blocks obtains higher quantitative measures on both datasets.

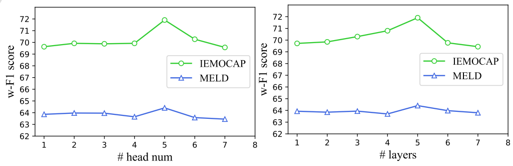


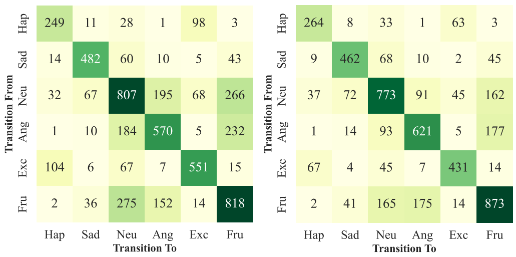
Figure 6: Effects of layer and head num used in MMTr on model.

6.6. Error Analysis

We study the IEMOCAP dataset in detail for error analysis of our model. We observe the emotion transition at the utterance level, i.e., emotion shift in adjacent utterances in the sampled utterance sample [Figure 7(a)] and speaker level, i.e., emotion shifts in utterances spoken by the same speaker [Figure 7(b)]. We find a high percentage of transitions between similar emotions, inferring that the models confuse similar classes of emotion. After analyzing the prediction results for the whole dataset, as shown in Figure 5(d), we predicted the happy category with a large proportion of its similar label excited, which verifies our

Journal Pre-proof

above speculation. We also found that the number of neutral categories transferred to other emotion categories was very high, resulting in more emotion categories being misclassified as neutral when classifying neutral emotion categories that were initially larger in sample numbers. This further reduces the effectiveness of the model classification.



(a) Utterance level (b) Speaker level

Figure 7: Utterance/Speaker-level Emotion transition in IEMOCAP. These are emotion transitions in consecutive utterances across/same speakers.

6.7. Case Study

Figure 8 shows a conversation sampled from the MELD dataset and a heatmap visualization of the three modal features of the current conversation. The conversation depicts a scenario in which Rachel and Ross, as a couple, argue because Rachel does not have time for Ross. In most cases, they feel sad or angry. At the beginning of the conversation, Ross's emotional state is neutral, while Rachel has already felt sad about Ross's words, so her emotional state shifts from the previous one. We can see from the heatmap that our feature representation of the text modal distinguishes between the first three sentences in which the emotional shift occurs. Over time, they become emotional. Both people are angry about each other's words. Although the text modal features at this point appear to be more similar in color than identical in conversations with the same Angry emotion, the final correct emotion classification is made by fusing other modal information through MMTr. The case appears not only as emotion shifting but also as a few emotion labels. As a more difficult dialog category to predict, MMTr can achieve correct prediction, indicating that selecting the correct main modal and then fusing it with other modal information can effectively alleviate the problem of emotion shifting and improve the correct rate of minority emotion category prediction.

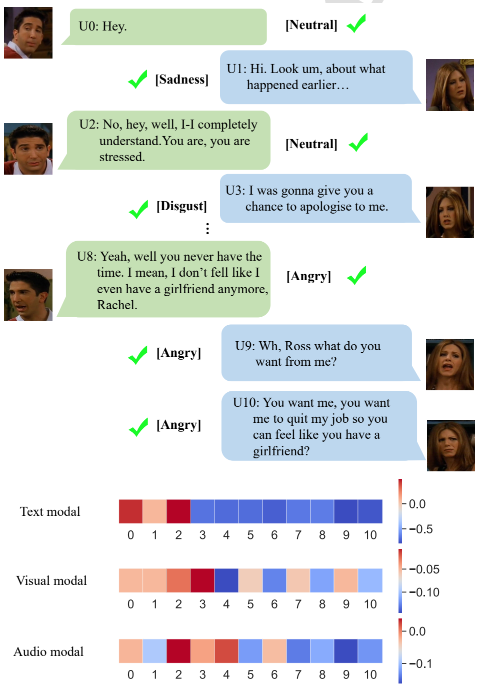

U0: Hey.
[Neutral] ✓
✓ [Sadness]
U1: Hi. Look um, about what happened earlier…
U2: No, hey, well, I-I completely understand.You are, you are stressed.
[Neutral] ✓
✓ [Disgust]
U3: I was gonna give you a chance to apologise to me.
⋮
U8: Yeah, well you never have the time. I mean, I don't fell like I even have a girlfriend anymore, Rachel.
[Angry] ✓
✓ [Angry]
U9: Wh, Ross what do you want from me?
✓ [Angry]
U10: You want me, you want me to quit my job so you can feel like you have a girlfriend?
Text modal
0 1 2 3 4 5 6 7 8 9 10
0.0
−0.5
Visual modal
0 1 2 3 4 5 6 7 8 9 10
−0.05
−0.10
Audio modal
0 1 2 3 4 5 6 7 8 9 10
0.0
−0.1

Figure 8: Case study in MELD.

7. Conclusion and future work

In this paper, we acquired emotional cues at both levels of the speaker's self-context and contextual context through an emotional cues extractor. In addition, we proposed a new multimodal fusion method that uses modalities with strong representational power as the main modal and enhances the representational power of weak modalities by preserving the integrity of their main modal features. We then designed a new multimodal ERC model, i.e., MMTr. We conducted comparative experiments on two benchmark datasets, and the experimental results showed that the model outperforms existing models with multimodal ERCs. In addition, the experimental results validated the correctness of our hypothesis of using text as the main modal.

However, MMTr still has shortcomings that must be improved in the future. For example, the two-level emotional cue extractor we use does not extract emotional cues well in multiperson conversations. Therefore we want to improve this method to adapt to multiperson conversation scenarios by extracting the corresponding features for each speaker. In addition, we currently use advanced feature extractors only for text modal, while for audio and visual modalities, we have not yet obtained good feature representations. Thus, in our future research, we need to extract more effective feature information of other modalities and improve the fusion strategy of the main modal transformer.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China [grant number 62141201], National Social Science Foundation of China [grant number

Journal Pre-proof

17XXW005], and Chongqing Ba'nan District Science and Technology Bureau[grant number 2020QC403]. We thank all the anonymous reviewers who generously contributed their time and efforts. Their professional recommendations have greatly enhanced the quality of the manuscript.

References

[1] J. Ni, V. Pandelea, T. Young, H. Zhou, E. Cambria, Hitkg: Towards goal-oriented conversations via multi-hierarchy learning, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 11112–11120.

[2] S. Yang, R. Zhang, S. Erfani, J. H. Lau, An interpretable neuro-symbolic reasoning framework for task-oriented dialogue generation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4918–4935. doi:10.18653/v1/2022.acl-long.338.

[3] T. Young, F. Xing, V. Pandelea, J. Ni, E. Cambria, Fusing task-oriented and open-domain dialogues in conversational agents, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 11622–11629.

[4] Y. Ma, K. L. Nguyen, F. Z. Xing, E. Cambria, A survey on empathetic dialogue systems, Information Fusion 64 (2020) 50–70.

[5] S. Shimojo, L. Shams, Sensory modalities are not separate modalities: plasticity and interactions, Current opinion in neurobiology 11 (2001) 505–509.

[6] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2247–2256.

[7] J. Hu, Y. Liu, J. Zhao, Q. Jin, 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), 2021, pp. 5666–5675.

[8] S. A. Gelman, S. O. Roberts, How language shapes the cultural inheritance of categories, Proceedings of the National Academy of Sciences 114 (2017) 7900–7907.

[9] M. A. Nowak, D. C. Krakauer, The evolution of language, Proceedings of the National Academy of Sciences 96 (1999) 8028–8033.

[10] M. Pagel, Human language as a culturally transmitted replicator, Nature Reviews Genetics 10 (2009) 405–415.

[11] K. A. Lindquist, M. Gendron, What's in a word? language constructs emotion perception, Emotion Review 5 (2013) 66–71.

[12] L. Devillers, L. Vidrascu, Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs, in: Ninth international conference on spoken language processing, 2006.

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536. doi:10.18653/v1/P19-1050.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.

[15] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 154–164.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 6818–6825.

[17] T. Ishiwatari, Y. Yasuda, T. Miyazaki, J. Goto, Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7360–7370.

[18] H. Ma, J. Wang, L. Qian, H. Lin, Han-regru: hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation, Neural Computing and Applications 33 (2021) 2685–2703.

[19] W. Li, W. Shao, S. Ji, E. Cambria, Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.

[20] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, Z. Yang, A multi-view network for real-time emotion recognition in conversations, Knowledge-Based Systems 236 (2022) 107751.

[21] L. Zhu, G. Pergola, L. Gui, D. Zhou, Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1571–1582.

[22] L. Yang, Y. Shen, Y. Mao, L. Cai, Hybrid curriculum learning for emotion recognition in conversation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11595–11603.

[23] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169–176.

[24] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2122–2132.

[25] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2594–2604.

[26] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6558–6569.

[27] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, L. Yi, Contrastive multimodal fusion with tupleinfonce, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 754–763.

[28] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9180–9192.

[29] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, D. Jiang, Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models, Knowledge-Based Systems 235 (2022) 107598.

[30] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 7037–7041.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter

Journal Pre-proof

Journal Pre-proof

of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT), 2019, pp. 4171–4186.

[33] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[34] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.

[35] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[36] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations(ICLR), 2015.

[37] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, 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), 2017, pp. 873–883.

[38] W. Shen, J. Chen, X. Quan, Z. Xie, Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13789–13797.

[39] D. Hu, L. Wei, X. Huai, 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), 2021, pp. 7042–7052.

Journal Pre-proof

Journal Pre-proof

Credit Author Statement

Shihao Zou: Conceptualization, Methodology, Investigation, Experiment, Software, Writing- original draft, Writing- Reviewing and Editing. **Xianying Huang**: Writing- Reviewing and Editing, Supervision. **Xudong Shen**: Writing- Reviewing and Editing, Visualization. **Hankai Liu**: Writing- Reviewing and Editing.

Journal Pre-proof

Journal Pre-proof

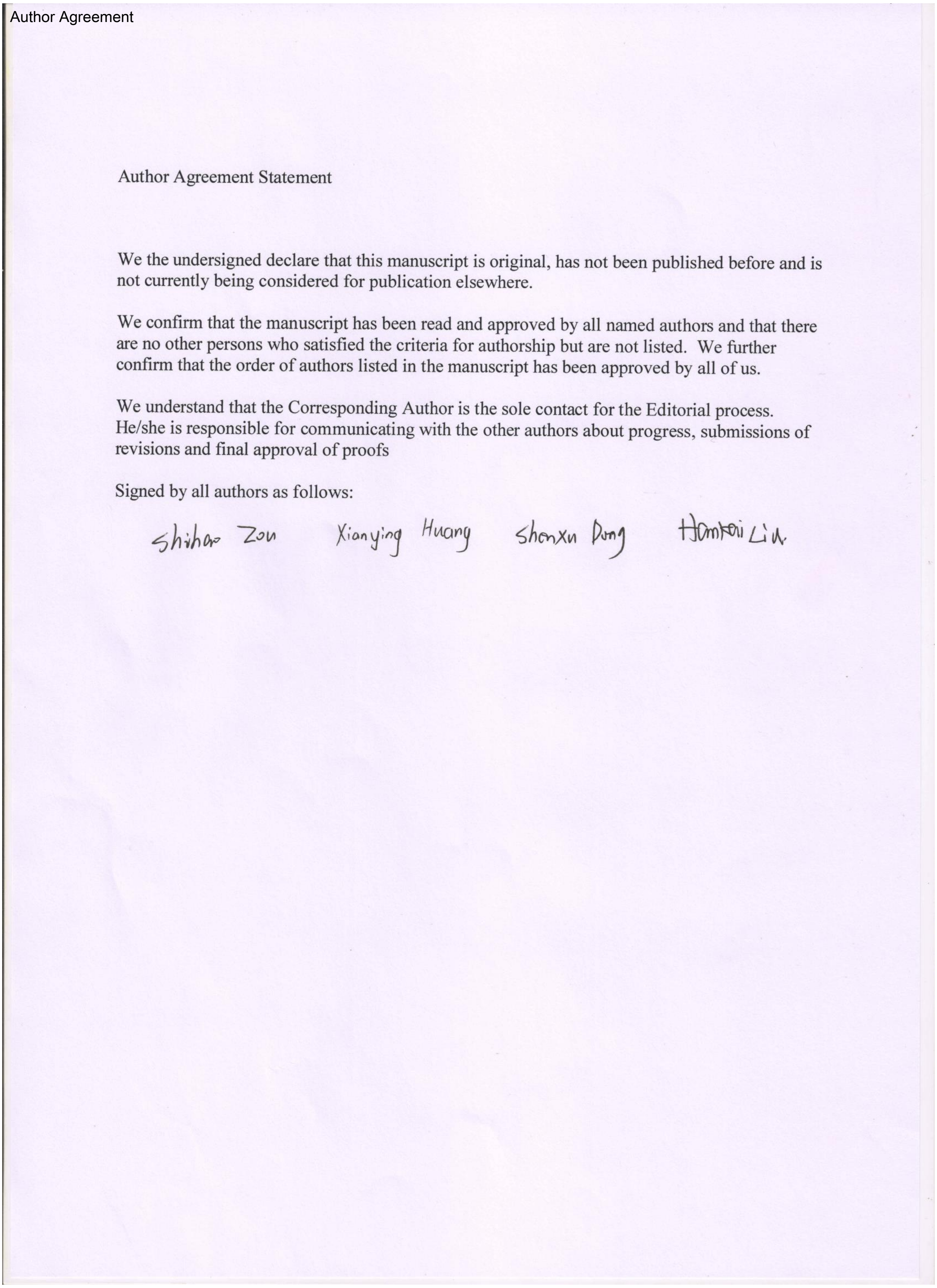
Title: Improving Multimodal Fusion with Main Modal Transformer for Emotion Recognition in Conversation

No.: KNOSYS-D-22-03529

Highlights:

1. Modal with different representational abilities should be learned differently.
2. Modal with stronger representation ability after feature extraction as the main modal.
3. Preserve the integrity of the main modal features, enhancing the weak modal feature.
4. Design an emotional cue extractor to enrich the conversation information.

Journal Pre-proof



Author Agreement

Author Agreement Statement

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs

Signed by all authors as follows:

Shihao Zou Xianying Huang Shenxu Dong Hankai Liu

Journal Pre-proof

Declaration of Interest Statement

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof