Intrinsic Emotion Information Enhanced the Causal Clause Recognition in Multimodal Conversations

Anonymous ACL submission

Abstract

Extracting emotions and their corresponding cause in multimodal conversation tasks, also known as MECPE has garnered considerable interest in recent years. Multimodal data contain intrinsic emotional information that is both associative and complementary, which can guide cause discovery and enable emotion-cause pairing. However, most researchers have overlooked the design of reliable mechanisms for effective guidance. Therefore, we propose Intrinsic Multimodal Emotion information with Masked Attention for Cause extraction(IMEMAC) to address this issue. To extract intrinsic emotion representations from multimodal information efficiently, we present a Two-Step Multimodal Fusion process that can fully integrate multimodal information. Then, the Masked Attention filter integrated information from different modalities to enhance the cause extraction performance. Finally, we leverage the recognized cause representation with intrinsic emotion representation to obtain emotion-cause pairs. The results of experiments demonstrate that the framework proposed in this paper achieves state-of-the-art performance in both the MECPE and MECPE-Cat tasks.

1 Introduction

Emotion-Cause Pair Extraction (ECPE) in conversations is one of the major challenge in the area of Natural Language Processing, which aims to extract emotion clauses in conversations and identify the corresponding causes of these emotions. Recognizing emotions and their corresponding causes in conversations is helpful for monitoring users' dialogue and analyzing emotional changes and potential reasons. Therefore, it is important that they have various emotional applications. (Gao et al., 2021) (Zhou et al., 2020).

Traditional ECPE focuses on textual conversation. However, human conversations in daily life inherently include multiple modalities *i.e.*, text, audio, video. With the proliferation of social media, people are increasingly engaging in online multimodal conversations. This has resulted in a wealth of multimodal data that provides a rich resource for research in multimodal emotion analysis. Consequently, research on extracting emotion and cause pair in multimodal conversation also known as the MECPE task has received significant interest in recent literature.

Current studies on ECPE mainly focus on collaboratively extract emotion and cause representations. For example, (Zhu et al., 2024) proposed a knowledge-guided graph attention network for ECPE, (Jiang et al., 2023) proposed Window Transformer for emotion and cause clause extraction, and (Xia and Ding, 2019) adapted a two-step solution to the ECPE tasks. In addition to text modality, (Wang et al., 2023a) firstly proposed the MECPE task and a new benchmark for this task. Multimodal data not only provide more information for conversation but also bring about more challenges. (Hu et al., 2024) adapted task-specific context aggregation for ERC and ECPE in multimodal settings. (Li et al., 2024) adopted holistic interaction and label constraint for the MECPE task. Despite their impressive performance, there remains a significant challenge: in practice, multimodal data contain associative and complementary intrinsic emotional information. Meanwhile, cognitive psychology research has suggested that emotions can be helpful in identifying causes(Weiner, 1985). This implies that because of the associative and complementary emotions within multimodal clauses clauses, considering their relationship can facilitate the pairing of causes and emotions. Although current methods involve emotional sentences in the cause extraction process(Li et al., 2024)(Hu et al., 2024), the corresponding causes are not associated with the emotions within multimodal data, even for the same causal clause, the intrinsic emotions among differ-

ent modalities can be different. Therefore, previous methods did not effectively inject intrinsic emotion information, which may cause the model to become confused when identifying causal clauses, resulting in suboptimal performance in pairing emotions with cause.

To tackle the issue mentioned above, this paper introduces Intrinsic Multimodal Emotion information with Masked Attention for Cause extraction (IMEMAC). Specifically, intrinsic emotion representations are injected into the cause-representation extraction process using a Masked Attention mechanism. Masked Attention mechanism was employed because different modalities within the multimodal clause have different effects on cause extraction. Our masked attention can filter integrated information from different modalities, thereby reducing confusion and enhancing the model's ability to detect causal clauses. To effectively extract these intrinsic emotion representations from multimodal information, a Two-Step Multimodal Fusion method is presented. First step combines information from textual, audio, and visual modalities in a pairwise manner, and the second step integrates these pairwise fusion results. Finally, we pair the emotional causes using the Cartesian product to obtain emotion-cause pairs. We carried out comprehensive experiments on benchmark datasets, and the results showed that our method achieved outstanding performance in both the MECPE and MECPE-Cat tasks. Additionally, various studies and analyses were performed to evaluate the effectiveness of each proposed module for MECPE tasks. The key contributions of this paper can be concluded as follows:

1) We propose Intrinsic Multimodal Emotion information with Masked Attention for Cause extraction to leverage emotion representation in assisting the extraction of causal representation.

2) We propose a Two-Step Multimodal Fusion method for intrinsic multimodal emotion extraction.

3) Extensive experiments demonstrate that the framework proposed in this study achieved appealing results in the MECPE task.

2 Related Works

2.1 Emotion Cause Extraction in Conversation

In recent years, increasing attention has been paid to emotion-related tasks, various tasks related to emotional reasons have emerged. (Poria et al., 2021) introduced the ECPE tasks, which aim to identify possible pairs of emotions and their associated causes in a document. (Chen et al., 2020) present an end-to-end method to capture the dependency information among local neighborhood candidate pairs. (Cheng et al., 2023) employ a two-turn machine reading comprehension task to address the ECPE task. (Zhao et al., 2023) adopted commonsense knowledge as bridges to address the problem of Causal Emotion Entailment. (Qian et al., 2023) represent the natural transition of emotion causes between two adjacent turns using an emotion cause transition graph.

2.2 Multimodal Learning in Conversation

Recently, there has been a surge in the analysis of multimodal conversational tasks. Multimodal analysis seeks to utilize various modal resources to improve the efficiency of specific downstream tasks. In multimodal Emotion Recognition in Classification (ERC), (Shi and Huang, 2023) adopted bidirectional multihead cross-attention layers to achieve multimodal fusion. (Zou et al., 2023) utilized filtered modality information as multimodal prompt information to fuse data from different modalities for ERC tasks. (Hu et al., 2022) employed intermodality contrastive learning to obtain comprehensive multimodal representations. In the field of multimodal sarcasm detection, (Wen et al., 2023) guide contrast learning by using label information to explore the potential interaction relationships among sarcastic expressions. (Zhang et al., 2021) introduced a multimodal metaphor benchmark to enhance the comprehension of metaphorical information from both multimodal text and images. (Wang et al., 2023b) developed a method for transforming emotion-cause-category triplets into a positional index sequence and employed pretrained models to generate this index sequence from input conversations. In the context of multimodal intent detection, (Zhang et al., 2022) proposed a multimodal intent detection benchmark that categorizes intent into two coarse-grained levels and 20 fine-grained levels. (Maharana et al., 2022) proposed a late fusion of multimodal embeddings to detect the intent of livestream videos effectively.

3 Problem Definition

A dialogue D consists of n utterances $D = \{u_1, ..., u_n\}$. In every dialogue, each utterance u_i

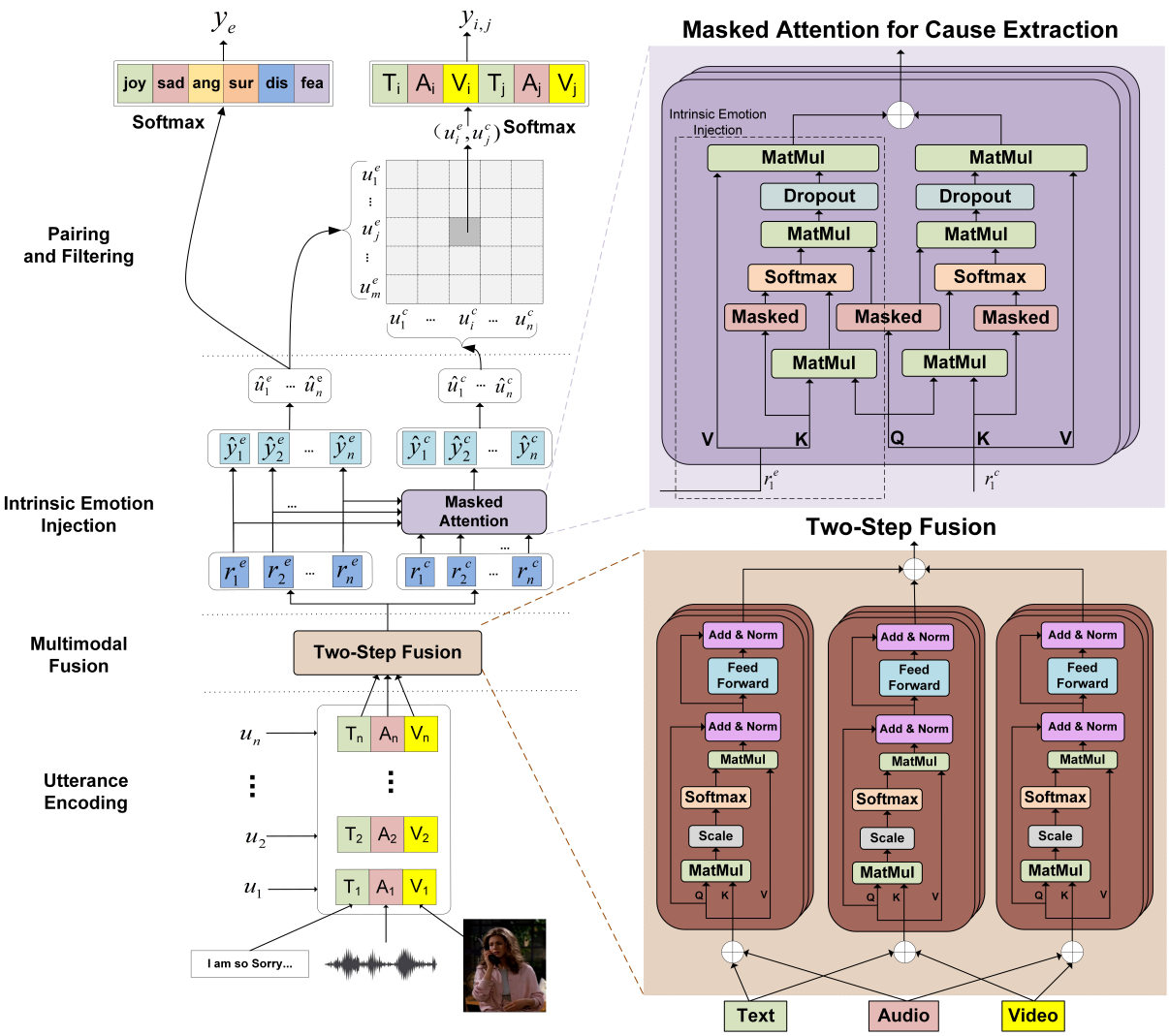

Figure 1: Illustration of the proposed method, from the bottom to the top: Utterance Encoding for preprocessing the multimodal data; Two-Step Fusion for multimodal fusion; Intrinsic Emotion Injection using extracted Intrinsic Emotion Representation and Masked Attention assist in Cause Extraction; Pairing and Filtering for extracting the emotion category of the emotion representation and pairing the multimodal emotion-cause pair

contains: textual information t_i, visual information v_i and audio information a_i, which can be presented as $u_i = [t_i, v_i, a_i]$.

For the MECPE task, given a dialogue D, the purpose is to extract the emotion-cause pairs $P = \left\{..., (u_i^e, u_j^c), ...\right\}$ from D based on information from textual, visual, and audio modalities. In the emotion-cause pair, P, u_i^e represents the extracted intrinsic emotion utterance and u_j^c represents the extracted cause utterance. Multimodal emotion–cause pair extraction with emotion category (MECPE-Cat) task requires the model to further recognize the specific emotion kinds of emotion-cause pair $P = \left\{..., (u_i^e, u_j^c, y^e), ...\right\}$, where y^e denotes the corresponding emotion category with $\left\{..., (u_i^e, u_j^e), ...\right\}$.

4 Method

4.1 The Overall Framework

The architecture of the proposed method is illustrated in Fig. 1. It is comprised of four primary components: Utterance Encoding, Multimodal Fusion, Intrinsic Emotion Injection for Cause extraction, Pairing and Filtering. Utterance Encoding involves the extraction of features from raw multimodal data. Multimodal Fusion integrates multimodal information using a two-step fusion process. Intrinsic Emotion Injection utilizes emotion data extracted from multimodal data to facilitate the cause-extraction process. Finally, the pairing

and filtering components were designed to obtain the multimodal emotion-cause pairs and specific emotion kinds.

4.2 Utterance Encoder

Utterance Encoder aims to convert the input multimodal dialogue into information that the subsequent modules can comprehend. Intuitively, textual modality is important for emotion recognition and reason extraction. To select a better encoder, we used Bidirectional Long Short-Term Memory network (BiLSTM) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) for textual features from extraction of the utterance because BiLSTM performs well on time sequence datasets, while BERT is known for its textual feature extraction ability. Before inputting the utterance to BiLSTM/BERT, a pre-trained 300-dimensional GloVe (Pennington et al., 2014) is adopted to initialize each utterance. In addition, we used OpenSMILE (Eyben et al., 2010) to extract the audio modality information and a convolutional 3D (Krause et al., 2021) network to extract the visual modality feature.

4.3 Two-Step Multimodal Fusion

After extracting features from the three different modalities, we designed a Two-Step Multimodal Fusion method. Early fusion aims to combine the primary feature-level information, is a widely adopted multi-modal fusion method. However, most early fusion methods concatenated all modalities directly into a new feature. This approach cannot facilitate in-depth feature interaction and fusion, which can lead to loss of interaction information between features.

In the proposed method, we concatenate different modalities in pairs to obtain the junior feature-level representation r_t^a, r_t^v and r_v^a.

$$r_t^a = r_t \oplus r_a \tag{1}$$

$$r_t^v = r_t \oplus r_v \tag{2}$$

$$r_v^a = r_v \oplus r_a \tag{3}$$

The operation mentioned above can better preserve the original modality information, allowing us to capture information between modalities that cannot be obtained by simply concatenating the three modalities. Subsequently, the junior fusion feature representation is passed to the multi-head attention mechanism to derive the corresponding feature representations. $\hat{r_t^a}$, $\hat{r_t^v}$ and $\hat{r_v^a}$. The multihead attention mechanism can consider the positions in the sequence simultaneously and effectively to establish long-distance dependencies.

$$\hat{r_t^a} = MultiHeadAttention(r_t^a) \tag{4}$$

$$\hat{r_t^v} = MultiHeadAttention(r_t^v) \tag{5}$$

$$\hat{r_a^v} = MultiHeadAttention(r_a^v) \tag{6}$$

Next, we further concatenate three fused features to obtain the final multimodal fusion result r_i

$$r_i = Concatenate(\hat{r_t^a}, \hat{r_t^v}, \hat{r_v^a}) \tag{7}$$

4.4 Intrinsic Emotion Injection for Cause Extraction

After acquiring the independent multimodal fusion representation r_i of utterance u_i, we input the r_i into the emotion extraction and the cause extraction encoders. Each encoder is based on either a BiLSTM or a Transformer, depending on whether BiLSTM or BERT is used to extract the textual modality features.

From a linguistic perspective, we intuitively consider that sentiment has the lowest level of abstraction, whereas emotion and sarcasm have the highest levels. In a multimodal setting, multimodal representations contain related and complementary internal emotional information. For example, visual modality information can complement emotion representation within textual modality. Meanwhile, cognitive psychology research suggests that in human thought processes, after understanding an emotion, people often look back to finding the causes (Clore and Bar-Anan, 2007), which means that emotion plays a guiding role in the process of cause discovery. It can be seen that to endow the model with cause extraction ability, we can design the model according to this principle. However, the corresponding causes are not associated with the emotions within multimodal data, which results in ignorance of the utilization of intrinsic emotion for cause detection.

Therefore, r_i^e from emotion extraction encoder can be regarded as the intrinsic emotion representation of the utterance, r_i^c can be viewed as the cause-specific representation of the utterance U_i. Then intrinsic emotion representation r_i^e and cause-specific representation r_i^c are fed into two $softmax$ layers to detect whether U_i is an emotion/cause utterance.

$$\hat{y_i^e} = softmax(W^e(r_i^e) + b^e)) \tag{8}$$

In our model, the results of multimodal emotion recognition not only perform the basic emotion recognition task but will be integrated into the cause discovery process to guide the identification of causes using the intrinsic emotions from multimodal data by Masked Attention Mechanism.

$$\hat{r_i^c} = MaskedAttention(r_i^e, r_i^c) \tag{9}$$

In the process of Masked Attention, we use intrinsic emotion representation to assist in extracting the cause representation, so r_i^c plays an important role. In the attention mechanism, the input acting as Q plays a dominant role; therefore, the input of Masked Attention will be processed as follows:

$$\begin{aligned} Q &= W^Q r_i^c \\ K_c &= W^{K_c} r_i^c \\ V_c &= W^{V_c} r_i^c \\ K_e &= W^{K_e} r_i^e \\ V_e &= W^{V_e} r_i^e \end{aligned} \tag{10}$$

Then Q and K multiply as normal attention mechanism to calculate the correlation and K will be masked and apply softmax on the masked outputs to make sure outputs can ignore some useless information.

$$output_1 = Softmax(QK_c) \tag{11}$$

$$output_2 = Softmax(QK_e) \tag{12}$$

Then the similar process will process on Q, the $Masked(Q)$ will perform another mask operation on the $outputs$

$$output_1^* = output_1 \times Masked(Q) \tag{13}$$

$$output_2^* = output_2 \times Masked(Q) \tag{14}$$

The outputs will go through a dropout layer and have a matrix multiply process with V as a common attention mechanism to determine how important each element is to the outputs.

$$output_c^{**} = Dropout(output_1^*) \times V_c \tag{15}$$

$$output_e^{**} = Dropout(output_2^*) \times V_e \tag{16}$$

Finally the $output_c^{**}$ and $output_e^{**}$ will be concatenated to get the final cause representation.

$$\hat{r_i^c} = output_c^{**} \oplus output_e^{**} \tag{17}$$

Then, the model uses a softmax layer to determine whether the cause representation extracted with the assistance of intrinsic emotions is a cause clause or not.

$$\hat{y_n^c} = softmax(W^c(\hat{r_i^c} + b^c)) \tag{18}$$

4.5 Pairing and Filtering

After extracting the intrinsic emotion utterances E and the cause utterance C from a dialogue, the model needs to pair these emotions with causes through casual relationships. First, we input E and C into the Cartesian product to obtain a set of emotion-cause pairs(u_e, u_c). Next, we need to identify whether the pair (u_e, u_c) is a valid pair or not. We use a feature vector x which entails the multimodal representations and distance information between two utterances to represent the pair and feed x into a softmax layer to filter the invalid pairs. For the MECPE-Cat task, the emotion representation will be fed into a softmax layer to determine which category the emotion representation belongs to.

$$\hat{y}_{j,k} = softmax(Wx_{(\hat{u_j^e}, \hat{u_k^c})} + b) \tag{19}$$

5 Experiments

In this section, we first introduce the datasets, followed by the baselines and the experiment settings. Then we will illustrate the result of MECPE and MECPE-Cat tasks. Ablation studies of the modules and case study about the model will presented subsequently

5.1 Datasets

We perform experiments on the ECF benchmark datasets (Wang et al., 2023a), which is obtained by annotating the causes of the emotion data in MELD (Poria et al., 2018).

5.2 Baseline

Because ECF is a new released data set with high difficulties. The main baselines are selected as follows.

MECPE-2Steps (Wang et al., 2023a): The baseline method proposed by MECPE data set, which detecting emotions and cause for the first step and then extracting emotion-cause pairs for the second step.

UniMEEC (Hu et al., 2024): UniMEEC share the prompt learning among modalities and then use

Methods		Emotion Extraction			Cause Extraction			Pair Extraction		
		P	R	F_1	P	R	F_1	P	R	F_1
MECPE-2steps	Text(BERT)	77.17	81.36	79.10	67.47	73.19	70.13	57.64	48.72	52.71
	+Audio	76.91	81.68	79.17	67.25	73.91	70.27	57.13	50.34	53.48
	+Video	77.10	81.18	79.03	6810	72.51	72.18	57.77	49.53	53.21
	+Audio+Video	0.7745	0.8110	0.7916	**68.42**	72.43	70.27	57.47	49.81	53.20
UNIMEEC	+Audio+Video	\	\	\	59.87	58.85	59.18	49.88	59.29	54.61
LLMs	GPT-3.5	57.75	65.08	61.19	39.87	29.30	33.78	8.95	51.23	15.24
	GPT-4	69.28	75.44	72.23	51.44	36.74	42.86	12.56	52.34	20.26
HiLo	Text	74.53	81.80	77.99	65.26	80.23	71.97	51.91	54.33	53.09
	+Audio	**78.31**	77.71	78.01	66.32	79.17	72.18	51.50	56.65	53.95
	+Video	76.03	80.14	78.03	61.98	**85.98**	72.01	54.60	52.71	53.64
	+Audio+Video	75.46	**83.66**	**79.35**	65.72	80.31	**72.28**	51.78	**59.69**	55.45
Ours	Text(BiLSTM)	75.07	79.43	77.14	63.47	78.64	70.19	58.76	50.39	54.16
	+Audio	74.74	79.56	77.01	63.93	76.86	69.67	58.29	51.59	54.60
	+Video	74.01	80.51	77.06	63.73	76.62	69.52	58.20	52.07	54.88
	+Audio+Video	74.49	80.17	77.18	63.77	77.43	69.88	57.69	53.77	**55.56**
	Text(BERT)	76.76	81.20	78.9	61.75	75.00	67.37	59.11	49.17	53.51
	+Audio	76.56	81.78	79.01	59.37	76.06	66.53	**59.67**	48.69	53.51
	+Video	76.87	81.53	79.08	60.92	75.45	67.20	58.61	50.32	54.03
	+Audio+Video	76.71	81.66	79.04	59.24	77.15	66.91	58.48	50.58	54.09

Table 1: Experiment Results on MECPE. SOTA is shown in **bold**, and the second place is underlined

aggregation from hierarchical context to manage the information of the task.

LLM-based Zero Shot: Using specific prompt on GPT-4 and GPT-3.5 (Ouyang et al., 2022) for zero-shot emotion-cause pair extraction.

HiLo (Li et al., 2024): HiLo adopted holistic interaction and label constraint and used emotion-switching features as key cues to enhance casual extraction in the conversation for the MECPE task.

5.3 Experiment Setting

In the setting of the data set, we divided the dataset into training, validation, and test sets in the ratio of 7:1:2 according to the steps of the benchmark model. The hidden dimension of BiLSTM is set to 100 and batch size is set to 32. All models were trained by Adam optimizer, learning rate set to 0.005, dropout set to 0.5, and L2 regularization weight set to 1e-5. For BERT, we adopt Bert-based model3, and the learning rate adopts linear pre-heating and linear attenuation mechanisms. Due to computational power limitations, we set the batch size of BERT to 4 and initial learning rate to 1e-5.

5.4 Results of MECPE

The experiment result on the emotion-cause pairing task is shown in Table 1. We use F_1 score as main evaluation metrics. Firstly, the F_1 score of our method on the MECPE task are better than the baseline methods, which obtain a score of 55.56%. Due to the computational power limitations, we cannot use the same parameters as BiLSTM on BERT to finish the MECPE Task, but experimental results on BERT still achieve a F_1 score of 54.09%, even the lowest F_1 score of 53.51% is better then MECPE-2steps on BERT, which shows better performance on BERT is to be expected, also shows that we inject intrinsic emotion representation into the detection process of the cause utterance by masked attention mechanism do increase the sensitivity of cause detection of intrinsic emotions, further enhancing the performance of extracting related emotion-cause pairs.

5.5 Results of MECPE-Cat

The experiment results on the emotion categories recognition task are shown in Table 2. Performance of our method on emotion kinds of surprise is the best and performance on Disgust is the worst. So the evaluation metric includes weighted-average F_1 score of without Disgust and Fear.

Similar to the experiment result on the MECPE task, the performance of the method we proposed is better than the baseline methods. The method we

Methods		Pair Extraction							
		Anger	Disgust	Fear	Joy	Sadness	Surprise	w-avg.6	w-avg.4
MECPE-2steps	Text(BERT)	24.39	00.00	00.71	38.84	21.60	40.24	29.32	31.92
	+Audio	26.30	00.46	00.50	39.62	22.73	41.35	30.47	33.15
	+Video	22.68	00.19	00.36	38.55	20.58	39.99	28.57	31.11
	+Audio+Video	25.08	00.09	01.01	38.49	22.36	40.64	29.62	32.24
HiLo	Text	28.95	8.82	14.89	39.75	19.87	44.15	32.14	34.05
	+Audio	28.85	1.77	3.51	41.38	23.50	43.67	32.39	35.09
	+Video	30.02	13.07	5.13	41.6	26.92	41.78	33.61	35.68
	+Audio+Video	29.17	1.68	3.12	42.58	25.10	43.55	33.04	**35.81**
Ours	Text(BiLSTM)	21.77	00.00	00.00	39.15	17.98	39.23	35.63	34.57
	+Audio	20.50	00.00	00.00	39.60	18.31	39.41	35.22	35.52
	+Video	21.06	00.00	00.00	39.14	18.58	39.02	34.72	34.72
	+Audio+Video	20.52	00.00	00.00	38.00	17.20	38.81	34.82	34.69
	Text(BERT)	21.45	00.17	00.53	39.31	25.30	40.27	34.72	34.50
	+Audio	22.66	00.00	00.83	39.52	26.41	40.35	35.66	35.13
	+Video	21.89	00.00	01.01	39.00	26.78	41.84	34.94	34.11
	+Audio+Video	22.04	00.36	00.53	39.59	25.13	40.72	**35.79**	35.43

Table 2: Experiment Result on MECPE-Cat. Representation is the same as Table 1.

TSF	IMEMAC	ECPE F_1
w/o	w/o	52.2
w/o	w/	53.93
w/	w/o	53.71
w/	w/	**55.56**

Table 3: Ablation study on MECPE

TSF	Text encoder	w-avg.6	w-avg.4
w/o	BiLSTM	28.72	30.98
w/	BiLSTM	34.82	34.69
w/o	BERT	29.62	32.24
w/	BERT	**35.79**	**35.43**

Table 4: Ablation study on MECPE-Cat

proposed achieves a weighted-average F_1 score on six emotion categories of 35.79% and, a weighted-average F_1 score on four main emotion categories of 35.52%. The improvement shows that the multimodal fusion method we proposed is helpful on the MECPE-Cat task.

The relatively poor performance on this task suggests that further predictions based on MECPE's emotional categories are more difficult.

5.6 Ablation experiment

Ablation experiments are conducted to demonstrate the effectiveness of the proposed modules. As shown in Table 3 and Table 4, by removing the Multimodal Fusion module, the experimental results of MECPE decrease from 55.56% to 53.95% in F_1 score, the results of MECPE-Cat decrease slightly in weighted average F_1 score of 6 emotions and 4 emotions. By removing the IMEMAC, the result of MECPE decreased from 55.56% to 53.71% on the F_1 score. By removing the Multimodal Fusion module and Masked Attention module, the result decreased slightly from 55.56% to 52.20%, which shows the effectiveness of the Two-Step Fusion and the IMEMAC.

To explore the impact of intrinsic emotion injection from different modalities on multimodal cause extraction, we injected emotion representation extracted from various modalities into the process of multimodal cause extraction. As shown in Table 5, multimodal emotion injection indeed improves the performance of cause entailment and the performance of emotion-cause pair recognition.

5.7 Parameter sensitivity analysis

We also evaluated the effect of parameters on the experimental results. We focus on the learning rate and the keep probability of audio, and visual modality. Keep_prob of 0 means the dropout of audio and visual modality is 1. The Results are shown in Fig. 2 and Fig. 3. We can see that the learning rate has a great impact on the results. The impact of the keep_prob also shows the importance of audio and visual modality in the process of MECPE.

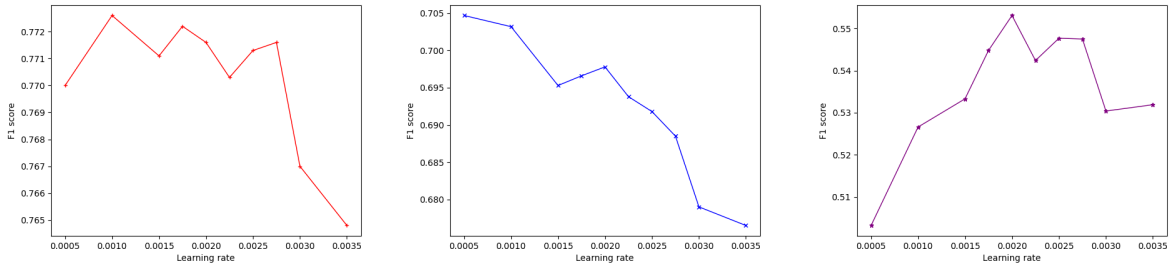

Figure 2: Impact of learning rate on F_1 score, from left to right are impacts on emotion recognition, cause recognition and emotion-cause extraction

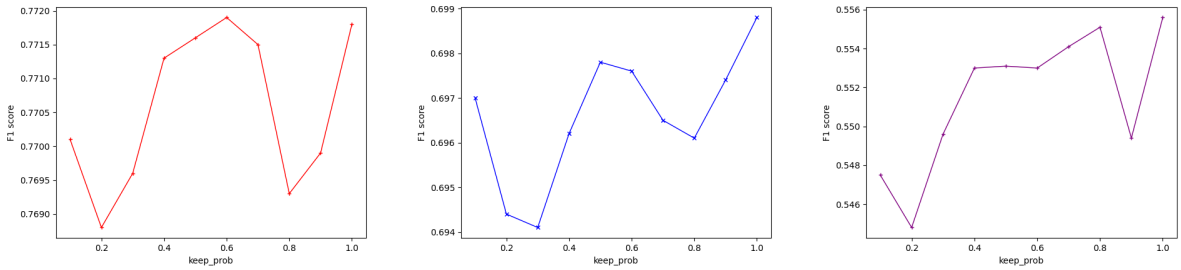

Figure 3: Impact of keep probability of audio and visual modalities on F_1 score, from left to right are impacts on emotion recognition, cause recognition and emotion-cause extraction

Injected modalities	CE F_1	MECPE F_1
text	68.48	53.45
text+audio	68.34	54.04
text+video	68.13	53.94
text+audio+video	**69.88**	**55.56**

Table 5: Ablation study on different injected modalities

5.8 Case Analysis

Furthermore, we analyzed some of the cases in our experiment to discuss the methods we proposed. The emotion recognition is the first step of the whole multimodal emotion-pair extraction task, so the result of the emotion recognition is worth analysis. In the data set, the 79th dialogue, when Mrs. Bing says "Are we okay now?", only uses BERT to extract the feature and then uses it to recognize the emotion failing to recognize the fear. However, the multimodal way to successfully recognize the emotion of the sentence is fear, because the voice modal supports more features of the sentence, which proves that multimodal fusion is helpful in emotion recognition, so as to cause utterance recognition. Meanwhile, our method achieves SOTA in matching emotion-cause pairs, which shows that our method is suitable for short conversation.

6 Conclusion

Based on the influence of emotion on human abductive cognition in cognitive psychology, considering the effective utilization of intrinsic emotional information to guide cause extraction, further achieving emotion-cause pair extraction, this paper proposes an Intrinsic Multimodal Emotion information with Masked Attention for Cause extraction (IMEMAC) method. We use Two-Step Multimodal Fusion to extract intrinsic emotion representation by firstly combining modalities information in pairs and secondly concatenating the result of the first step. Then we inject Intrinsic Emotion representation extracted by Two-Step Multimodal Fusion into Cause extraction with Masked Attention mechanism. Masked Attention can filter integrated information from different modalities, thereby reducing confusion and enhancing the ability to distinguish causal clauses. Finally, the Cartesian product is adopted to pair the intrinsic emotion representations and the cause representations. Experimental results show that our method achieves the SOTA on the MECPE process. We believe this research can provide broader insights into how machines can understand human emotions and causes.

Limitations

Despite the good performance achieved by our proposed method for the MECPE task, there are still certain limitations. First, our method injects intrinsic emotion information into the extraction process of reasoning representation without considering the injection of contextual emotional representation. Therefore, our method may perform less effectively when addressing emotion representation and cause representation that is located farther away compared to those that are closer. Additionally, our approach only considers attention-based methods for emotional injection, without exploring the effects of other injection techniques. In the future, we will concentrate on the impact of emotional injection at varying distances on reasoning extraction, which explores the ability to process long multimodal conversations; simultaneously, we will also explore the effects of graph-based or other mechanisms with commonsense knowledge from visual and audio modalities injection on MECPE.

Acknowledgments

The authors would like to express their appreciation and gratitude to all reviewers for their constructive and helpful feedback.

References

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, Shiping Ge, and Qing Gu. 2023. A consistent dual-mrc framework for emotion-cause pair extraction. *ACM Trans. Inf. Syst.*, 41(4).

Gerald Clore and Y. Bar-Anan. 2007. Affect-as-information. *Encyclopedia of Social Psychology*, pages 14–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yunlv Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Hasti Seifi, and Jiayuan Xie. 2024. Unimeec: Towards unified multimodal emotion recognition and emotion cause. *Preprint*, arXiv:2404.00403.

Dazhi Jiang, Hao-Ying Liu, Geng Tu, and Runguo Wei. 2023. Window transformer for dialogue document: a joint framework for causal emotion entailment. *International Journal of Machine Learning and Cybernetics*, 14:2697–2707.

Daniel Krause, Archontis Politis, and Konrad Kowalczyk. 2021. Comparison of convolution types in cnn-based feature extraction for sound source localization. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 820–824.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Trans. Multimedia Comput. Commun. Appl.* Just Accepted.

Adyasha Maharana, Quan Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ArXiv*, abs/1810.02508.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Ni Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13.

Yushan Qian, Bo Wang, Ting-En Lin, Yinhe Zheng, Ying Zhu, Dongming Zhao, Yuexian Hou, Yuchuan Wu, and Yongbin Li. 2023. Empathetic response generation via emotion cause transition graph. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. Multimodal emotion-cause pair extraction in conversations. *IEEE Trans. Affect. Comput.*, 14(3):1832–1844.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023b. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963, Singapore. Association for Computational Linguistics.

Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review*, 92 4:548–73.

Zhiyuan Wen, Rui Wang, Xuan Luo, Qianlong Wang, Bin Liang, Jiachen Du, Xiaoqi Yu, Lin Gui, and Ruifeng Xu. 2023. Multi-perspective contrastive learning framework guided by sememe knowledge and label information for sarcasm detection. *International Journal of Machine Learning and Cybernetics*, 14:4119 – 4134.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. 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 3214–3225, Online. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1688–1697, New York, NY, USA. Association for Computing Machinery.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Peican Zhu, Botao Wang, Keke Tang, Haifeng Zhang, Xiaodong Cui, and Zhen Wang. 2024. A knowledge-guided graph attention network for emotion-cause pair extraction. *Knowledge-Based Systems*, 286:111342.

Shihao Zou, Xianying Huang, and Xudong Shen. 2023. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5994–6003, New York, NY, USA. Association for Computing Machinery.