ORIGINAL ARTICLE

Dialogue emotion model based on local–global context encoder and commonsense knowledge fusion attention

Weilun Yu[1] · Chengming Li[2] · Xiping Hu[3] · Wenhua Zhu[1] · Erik Cambria[4] · Dazhi Jiang[1]

Received: 12 September 2023 / Accepted: 8 December 2023 / Published online: 9 January 2024

Abstract

Emotion Recognition in Conversation (ERC) is a task aimed at predicting the emotions conveyed by an utterance in a dialogue. It is common in ERC research to integrate intra-utterance, local contextual, and global contextual information to obtain the utterance vectors. However, there exist complex semantic dependencies among these factors, and failing to model these dependencies accurately can adversely affect the effectiveness of emotion recognition. Moreover, to enhance the semantic dependencies within the context, researchers commonly introduce external commonsense knowledge after modeling it. However, injecting commonsense knowledge into the model simply without considering its potential impact can introduce unexpected noise. To address these issues, we propose a dialogue emotion model based on local–global context encoder and commonsense knowledge fusion attention. The local–global context encoder effectively integrates the information of intra-utterance, local context, and global context to capture the semantic dependencies among them. To provide more accurate external commonsense information, we present a fusion module to filter the commonsense information through multi-head attention. Our proposed method has achieved competitive results on four datasets and exhibits advantages compared with mainstream models using commonsense knowledge.

Keywords Emotion recognition in conversation · Local–global encoder · Commonsense knowledge · Multihead attention

1 Introduction

As an emerging research direction in Natural Language Processing (NLP), Emotion Recognition in Conversation (ERC) is attracting more and more attention from researchers due to its important role in various emotional application areas [1, 2], such as opinion mining in social media [3–5], healthcare systems based on psychoanalytic tools, and emotionally intelligent and autonomous conversational robots [6–8]. Despite significant research achievements in recent years, the current models for ERC still need further improvement. The main reason is that most models lack the ability to connect the semantic information of context and speakers' psychological state when modeling utterances and speakers. The psychological factors of speakers can enrich the semantic information of dialogue. To optimize dialogue and speaker modeling using these latent factors, introducing an external commonsense knowledge base is necessary.

In recent studies on introducing the external commonsense knowledge base, Zhong et al. [9] used a graph attention network to dynamically perceive the emotional information of the context. Poria et al. [10] proposed COSMIC, a

✉ Erik Cambria
erik@sentic.net

✉ Dazhi Jiang
dzjiang@stu.edu.cn

Weilun Yu
21wlyu@stu.edu.cn

Chengming Li
licm1130@gmail.com

Xiping Hu
huxp@bit.edu.cn

Wenhua Zhu
21whzhu@stu.edu.cn

[1] Department of Computer Science, Shantou University, Shantou, China

[2] Department of Electronic and Computer Engineering, Shenzhen MSU-BIT University, Shenzhen 518172, China

[3] School of Medical Technology, Beijing Institute of Technology, Beijing, China

[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

commonsense-oriented model built on a huge commonsense knowledge base. Li et al. [11] further proposed the SKAIG-ERC model, which chose five commonsense knowledge as the mental state of dialogue to construct a graph neural network. Although the models mentioned above have achieved reliable results, there is still room for improvement.

Research has shown that long-distance contextual information greatly enhances the prediction of speakers' emotions [12]. However, previous models have not effectively integrated intra-utterance information, local contextual information, and global contextual information in predicting the emotion of the current utterance. Given the complex semantic dependencies among these factors, obtaining utterance vectors that contain information from all three sources in conversation scenarios is an urgent challenge in the emotion recognition task (ERC). When people convey emotional information in daily conversations, in addition to the semantic information mentioned above, there is also the influence of commonsense knowledge [13]. However, most existing models only mechanically inject the extracted commonsense knowledge, rather than dynamically and organically integrating the commonsense knowledge into the dialog model. This can introduce unexpected noise and greatly reduce the efficacy of commonsense knowledge in ERC [14]. As in Fig. 1, we chose a clip from Friends to illustrate these issues.

To address these issues, this paper proposes an ERC model based on a local–global context encoder and commonsense knowledge fusion attention. Our model is divided into two main blocks: the local–global context encoder and the commonsense knowledge fusion modules effectively integrate semantic information inside the dialogue and commonsense knowledge outside the dialogue, respectively. They naturally combine the information inside and outside the conversation. In order to integrate the information within the dialogue, enabling the model to efficiently extract information from distant contexts, near contexts, and within utterances, we designed the Global-Local Context Encoder. Utterances are encoded by the local context encoder to obtain intra-utterance and local contextual information. We designed a global context encoder stacked on the global state to model the relative position of the utterance in the conversation, capture global context information, and address long-distance dependencies among the utterances. We then employ an attention block to process the information extracted by the local context encoder and the global context encoder to obtain a vector with intra-utterance, local contextual, and global contextual information. Commonsense knowledge is injected after local–global context encoding to help the model better understand context information. While commonsense knowledge enriches the context semantics, it inevitably introduces noise information, which interferes with the context enrichment effect. Therefore, a Commonsense Knowledge Fusion Attention module is proposed to integrate commonsense information and facilitate target selection. The commonsense knowledge fusion part first generates the inferential commonsense knowledge features of each utterance under the specific reasoning relationship through the commonsense knowledge generation model (COMET) [15]. Then, the different types of commonsense knowledge are put into a commonsense knowledge fusion

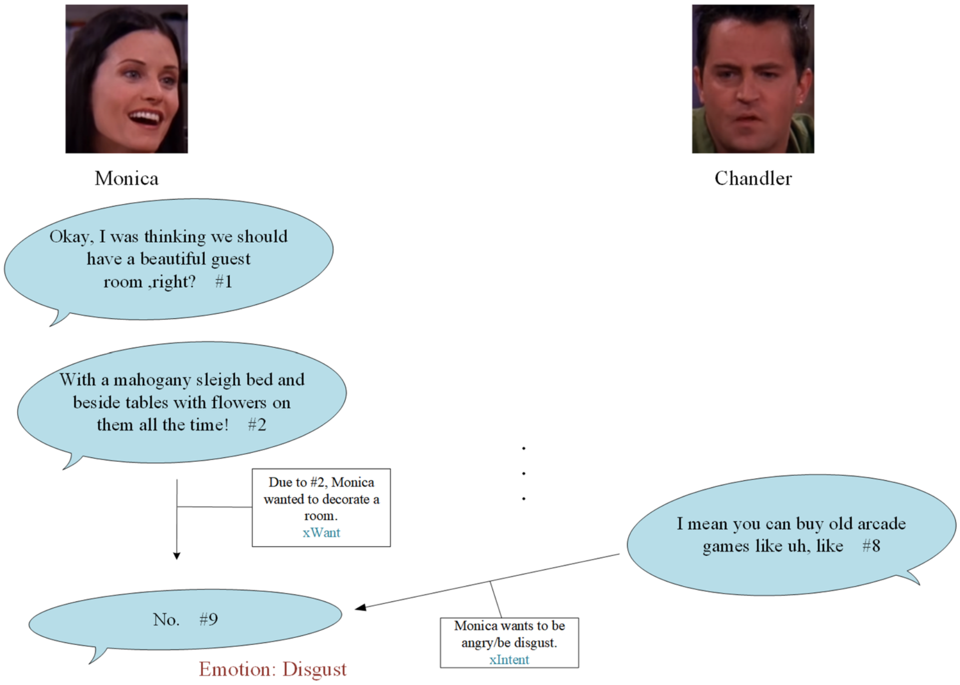
Monica
Chandler
Okay, I was thinking we should have a beautiful guest room .right? #1
With a mahogany sleigh bed and beside tables with flowers on them all the time! #2
Due to #2, Monica wanted to decorate a room.
xWant
. . .
I mean you can buy old arcade games like uh, like #8
No. #9
Emotion: Disgust
Monica wants to be angry/be disgust.
xIntent

Fig. 1 A conversation clip between Monica and Chandler in MELD dataset. The utterance #2 provides the xWant of Monica for #9, and #8 provides the intention. Both give positive and rational hints for #9 to predict the negative emotion disgust

module (State Attention), which helps the networks integrate commonsense information.

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

1. We propose a local–global context encoder that addresses the shortcoming of past encoders that ignore long-distance context information and effectively integrates information of intra-utterances, short-distance contexts as well as long-distance contexts. It not only can obtain richer semantic information at the dialogue encoder layer by achieving local and global context sensitivity, but also organically combine information outside the dialogue.
2. We propose a novel commonsense knowledge fusion module that improves the relevance of commonsense knowledge in conversations, unlike the past mechanical injection of commonsense knowledge into conversation models. Weighted selection of different types of commonsense knowledge can accurately capture the emotions embedded behind even in conversations with larger amounts of data and more complex interactions.
3. We evaluate our proposed model on four publicly available datasets (MELD, IEMOCAP, EmoryNLP, and DailyDialog) commonly used in ERC tasks. The results show that our model achieves competitive performance compared to benchmark models and outperforms other models that use commonsense knowledge.

The rest of this article is organized as follows: Sect. 2 reviews related work; Sect. 3 describes the proposed model, including the local–global context encoder, commonsense knowledge extraction and fusion, and emotion classification; Sect. 4 reports the experimental setup, including the experimental data, baseline methods, evaluation metrics, and parameter settings; Sect. 5 illustrates and analyzes the experimental results; finally, Sect. 6 concludes the article and proposes concluding remarks.

2 Related work

ERC has attracted significant attention from researchers, and several methods have been proposed to improve the accuracy of recognition, such as multi-modality [16, 17] and multi-task approaches [18]. In this section, we introduce research progress on contextual information modeling and the application of commonsense in this area.

2.1 Contextual information modeling

Contextual information has always been at the core of research in the field of Natural Language Processing (NLP), and recent studies [19–21] have shown that effective capturing of contextual information can significantly improve various NLP tasks. In dialogue data, the utterances near the target utterance can be used as contextual information [22]. Proper use of contextual information can help the model understand the content of the conversation and recognize the emotion of the current utterance in context.

Poria et al. [23] proposed the bi-directional contextual LSTM (bc-LSTM) model, which has demonstrated through experiments that contextual information can help the model make correct emotional predictions. The bidirectional LSTM module of the model takes advantage of the natural order of the recurrent neural network to consider the dialogue content as sequential data and effectively captures contextual information through neural units at different moments. Zahiri et al. [24] proposed a convolutional neural network model, which is modeled with sequential patterns for recognizing the sentiment of utterances in conversations. Additionally, the authors incorporated an attention mechanism between multiple layers of convolutions to capture different information learned by the layers. The aforementioned methods are based on the sequence structure to model contextual information. In recent years, the graph neural network model has achieved good results in many fields [25]. As a result, some researchers have applied the graph neural network model to the contextual information modeling of ERC. Zhang et al. [26] used a graph model to solve the task of dialogue emotion recognition involving multiple people. The authors considered each speaker and utterance as nodes, and the edges between nodes as context dependencies. These dependencies capture contextual information. Shen et al. [27] proposed a model for encoding utterances with a directed acyclic graph and a directed acyclic neural network. The model combines the advantages of the graph neural network model and the sequence model to simulate the information flow between the long-distance context and the short-distance context in a more intuitive way.

2.2 Commonsense knowledge

Although progress has been made in using context-assisted analysis of dialogue sentiment, limited information is available within the dialogue [28–30]. To address this limitation, researchers have proposed using additional information to assist models in understanding dialogue content [31], such as commonsense knowledge. With the development of commonsense knowledge bases [32, 33], some researchers have employed these bases for related tasks [34, 35]. Based on this, there are many works that try to combine context and commonsense knowledge in other tasks of NLP. Liu et al. [36] improves commonsense reasoning in text generation tasks by incorporating knowledge graphs into pre-trained models to produce more logical and natural sentences as output. Zhang et al. [37] fuses encoded representations from pre-trained models and graph neural networks through multi-layer modal interactions. Information from one modality is propagated to the other so that contextual representations are based on structured world knowledge and so that linguistic nuances in the context inform graphical representations of knowledge. Song et al. [38] introduces global and local knowledge constraints method which makes the pretrained model better adapt to the multilingual knowledge graph completion task. The former is used to constrain the reasoning of answer entities, while the latter is used to enhance the representation of query contexts.

In the field of ERC, Zhong et al. [9] were the first to use a commonsense knowledge base for emotion recognition and proposed an ERC model based on external knowledge enhancement. The authors introduced external commonsense knowledge into the model and utilized graph attention to make commonsense knowledge dynamic. This method incorporates emotional information of the perceptual context to give commonsense knowledge emotional characteristics. Ghosal et al. [10] used external commonsense knowledge to help the model simulate changes in the speaker's mental state, intention, and emotional state during the dialogue. Their objective was to address the difficulties in contextual information dissemination and emotion transition monitoring in ERC tasks. Zhu et al. [39] proposed a Transformer model for topic-driven and knowledge-aware dialogue analysis. They believe that identifying the topic in the dialogue is crucial to understanding the content of the conversation. The authors integrated commonsense knowledge into the utterance encoder information through the attention mechanism, allowing the utterances to encode the ability to perceive the topic of the conversation. In order to explore the potential value of commonsense knowledge in ERC, Tu et al. [40] proposed Sentic GAT, which tends to select commonsense knowledge that correlates with the contextual semantics of the target discourse and with the degree of sentiment. Based on this, Jiang et al. [41] proposed a GESM approach with sentiment consistency to reduce the size of commonsense knowledge. It tends to search for knowledge that is emotionally consistent with the vocabulary. After that, a genetic algorithm is added to constitute LESM, which is capable of selecting the knowledge obtained by GESM, thus effectively improving the quality of external knowledge. In particular, it can compensate for its limitation of selecting knowledge in negative contexts.

3 Methodology

3.1 Task definition

Given information about the dialogue and the corresponding speaker, the ERC task aims to identify the latent sentiment contained in each utterance in a dialogue from a set of predefined sentiment categories, where each utterance has a human-labeled sentiment label, such as sadness, happiness, fear, etc. Formally speaking, given an input sequence consisting of N utterances $\{(u_1, p_1), (u_2, p_2), ..., (u_N, p_N)\}$, where each utterance $u_i = \{u_{i,1}, u_{i,2}, ..., u_{i,T}\}$ is made up of T words and spoken by speaker p_i. The ERC task requires predicting the emotional label e_i of each utterance u_i.

3.2 Overview of dialogue emotion model

We propose a dialogue emotion model for the ERC task that combines a local–global context encoder and commonsense knowledge fusion attention. Figure 2 shows the architecture of the proposed model. The model consists of two main parts: a local–global context encoder module and a commonsense knowledge fusion module. In the local–global context encoder, we extract utterance-level features and inter-utterance location information to capture the contextual information of the dialogue. The extracted semantic information is then input into the commonsense knowledge fusion module, which incorporates external commonsense knowledge extracted by COMET. The output and the commonsense state at the previous time step are further processed using a gated recurrent unit (GRU) [42] to obtain three commonsense states related to the psychological state at the current moment: the internal state, external state, and intent state. These three states are then input to the commonsense knowledge fusion module (State Attention) to obtain a new state that better describes the psychological situation of the speaker. The proposed model leverages contextual information from the entire conversation to predict the correct sentiment label for each utterance. In the following subsections, we provide a detailed description of the model design.

3.2.1 Local–global context encoder

3.2.1.1 Local context encoder For each utterance in a dialogue, it is necessary to extract utterance-level information about the intra-utterance and local context from its words. To this end, we adopt the modified RoBERTa model, which is similar to the BERT Large [43] architecture, proposed by Kim et al. [44] as the local context encoder. We treat each utterance as a sequence of tokens with an emotion label. In this setting, the fine-tuning of the pretrained RoBERTa model is performed on the task of utterance classification. Specifically, for n utterances $\{u_1, u_2, \ldots, u_n\}$ in a dialogue, we insert a token *CLS* before each utterance to obtain a sequence $\{cls_1, u_1, cls_2, u_2, ..., cls_n, u_n\}$, which was inspired by Liu and Lapata [45].

The processed sequence is then classified into its sentiment label in a small feed-forward network using the activation from the last layer corresponding to the *CLS*

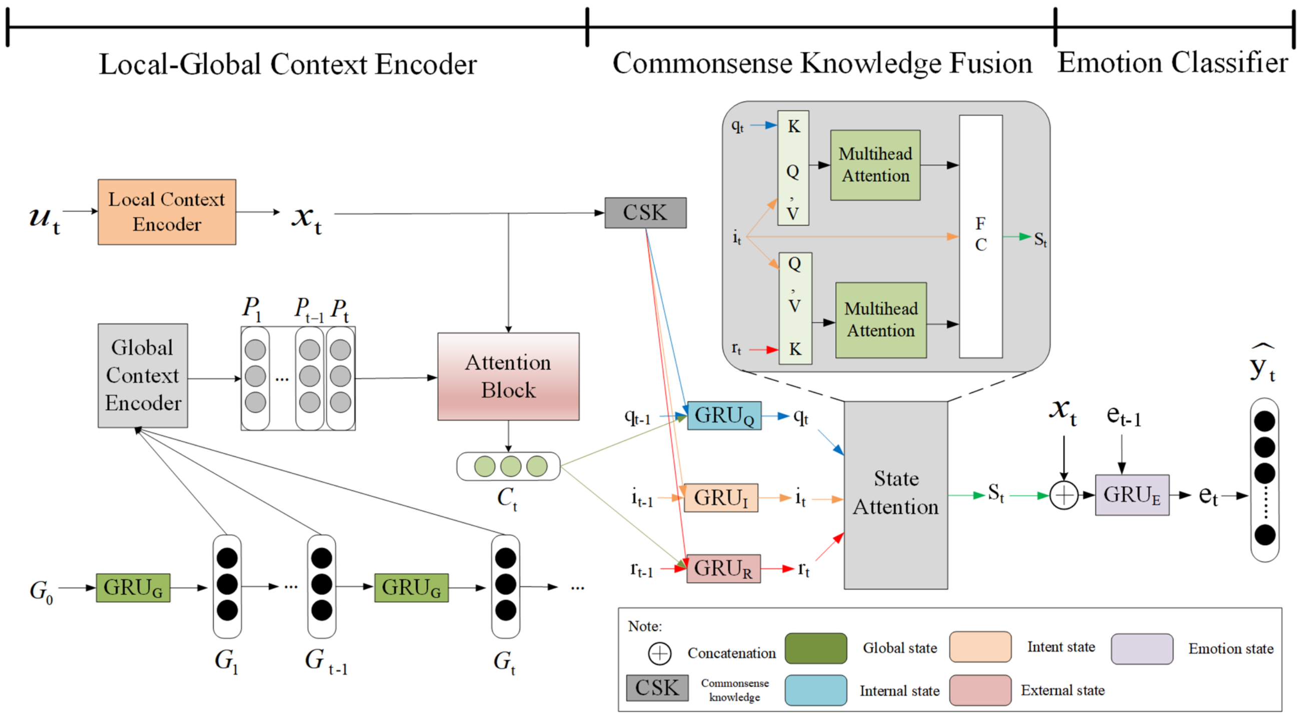

Fig. 2 Local–global context encoder and commonsense knowledge fusion attention model. CSK means the commonsense knowledge from COMET. State Attention is the commonsense knowledge fusion module. FC indicates the fully connected layer

token. In RoBERTa, the sequence $\{cls_1, u_1, cls_2, u_2, \ldots, cls_n, u_n\}$ is used as input, and a sequence $\{x_{cls_1}, x_1, x_{cls_2}, x_2, \ldots, x_{cls_n}, x_n\}$ is output.

$$x_t = RoBERTa(u_t),\ t \in [1, n],\ \forall x_t \in R^n \tag{1}$$

Each utterance u_t is represented by a vector $x_t \in R^n$. We use these vectors as representations of the utterances, which are then accompanied by the text of their previous and next utterances.

3.2.1.2 Global context encoder *Global state* The purpose of global state is to jointly encode the utterance and commonsense state to capture a given context. These states help analyze the speaker's psychological state, resulting in improved contextual representation. The currently encoded utterance x_t changes the speaker's psychological state from the immediate previous time-step, capturing this change using GRU_G. The output of the global state is denoted as $\{G_1, G_2, \ldots, G_n\}$.

$$G_t = GRU(G_{t-1}, x_t \oplus r_{s(u_t),t-1} \oplus q_{s(u_t),t-1}). \tag{2}$$

Global context encoder We calculate the relative spatial separation between the global states at the target time and the candidate time, leveraging this positional information to enhance the representation of the utterance.

$$P_i = G_i \oplus pemb_i,\ i \in [1, t-1]. \tag{3}$$

Context state The context state, captured by combining local and global contexts, stores rich dialogue information. In Eq. 5, attention scores α_i are calculated over the global context encoder with relative position embedding P_i and the output of local context encoder x_i. This results in higher attention scores for utterances that have a greater impact on dialogue emotions. Finally, in Eq. 6 we pool the attention vector a_t from the history of the global context encoder with the relative position embedding $\{p_1, p_2, \ldots, p_{t-1}\}$.

$$a_i = tanh(W_s p_i + b_s),\ i \in [1, t-1] \tag{4}$$

$$\alpha_i = \left(\left(a_i^{\mathrm{T}} x_i\right)^e \Big/ \sum_{i=1}^{t-1} (a_i^{\mathrm{T}} x_i)^e \right) \tag{5}$$

$$c_t = \sum_{i=1}^{t-1} \alpha_i P_i. \tag{6}$$

3.2.2 Commonsense knowledge fusion

3.2.2.1 Commonsense knowledge feature extract In this stage, we utilize the Commonsense Transformer model (COMET) to extract features of commonsense knowledge. COMET utilizes the pretrained autoregressive language model GPT [46] as its foundation and is equipped with the ability to automatically construct knowledge bases through training on multiple commonsense knowledge graphs.

COMET is trained on the Atlas of Machine Commonsense (ATOMIC) [33], which is a compilation of everyday if-then commonsense knowledge presented in textual descriptions, to generate commonsense knowledge structures. ATOMIC focuses on commonsense knowledge organized in typed if-then reasoning relations, such as “if *X* makes *Y*’s coffee, then *Y* will be appreciative.” ATOMIC focuses on event sequences and related social commonsense knowledge, and is a knowledge base for commonsense reasoning about daily events. It defines relationships for nine types of reasoning: *x*Intent, *x*React, *x*Want, *x*Effect, *x*Attr, *x*Need, *o*Want, *o*React, and *o*Effect. Therefore, it can be formalized as a triple s, r, o, where s is the subject, r is the relation, and o is the object. For example, several tuples related to the event “X makes Y’s coffee” can be: (s = “if *X* makes *Y*’s coffee”, r = “*x*Intent”, o = “*X* wanted to be helpful”), (s = “if *X* makes *Y*’s coffee”, r = “*o*React”, o = “*Y* will be appreciative”), and (s = “if *X* makes *Y*’s coffee”, r = “*x*Effect”, o = “*X* will get thanked”).

Inspired by Ghosal [10], although COMET is an encoder-decoder model, we only use the encoder part and discard the decoder. To process the input, we concatenate the subject phrase U with the relation phrase r and then feed the resulting sequence $U \oplus r$ through COMET’s encoder, extracting activations from the final time-step, denoted as CSK. The intention behind this is that r is represented in the form of a discrete vector after being encoded by the encoder. From the perspective of vector space, it is similar to a word vector representation and has the ability to represent semantic information. COMET can generate commonsense knowledge with main event reasoning relations through the decoder because, after being trained on the basis of the ATOMIC knowledge base, the vector representation of the relation r has certain reasoning semantics. Therefore, after the pre-trained encoder part of a given relation r, the generated vector representation is necessary for the subsequent modeling of contextual information.

At a more detailed level, Ghosal [10] believes that the speaker’s intent state, internal state, and external state are all important for understanding the essence of dialogue. Therefore, we choose to jointly model these three states to understand the speaker’s mental state. Emotion states are then modeled in terms of combinations of the three states and previous emotion states. These commonsense models are performed using Bidirectional GRU cells. The current hidden state h_t is updated by GRU cells, which take commonsense knowledge CSK, context state c_t, and the previous hidden state h_{t-1} as input, according to the following formula:

$$h_t = GRU(h_{t-1}, c_t \oplus CSK) \tag{7}$$

Three Bidirectional GRU cells, GRU_Q, GRU_R, and GRU_I, are used to model the internal state, external state, and intent state, respectively.

Internal state The speaker’s internal state is influenced by their emotions and the impacts they perceive from other speakers. Expressing their feelings or opinions is not always done through explicit stances, outward behaviors, or exaggerated responses. This state is covert because it includes aspects that the speaker may be unwilling to express or is considered a characteristic that does not need to be explicitly stated. In summary, the influence on the speaker itself is the basic factor that constitutes the speaker’s internal state. We use GRU_Q to model the speaker’s internal state.

$$q_{s(u_t),t} = GRU_Q(q_{s(u_t),t-1},\ c_t \oplus CSK(x_t)),\ t \in [1, n]. \tag{8}$$

External state In contrast to the internal state, the external state of the speaker is not hidden; it is all about expressing, reacting, and responding. This state is readily visible, perceptible, or comprehensible to other speakers. It encompasses a range of modalities, including visible expressions, the speaker’s speech patterns, acoustic features, as well as visual cues like gestures and postures, which collectively constitute the external characteristics of the speaker. We use GRU_R to model the speaker’s external state.

$$r_{s(u_t),t} = GRU_R(r_{s(u_t),t-1},\ c_t \oplus CSK(x_t)),\ t \in [1, n] \tag{9}$$

Intent state The intent state represents a commitment to perform a specific set of actions. According to Ghosal [10], the emotional state of a conversation is always significantly influenced by the speaker’s intention. We use GRU_I to model the speaker’s intent state.

$$i_{s(u_t),t} = GRU_I(i_{s(u_t),t-1},\ CSK(x_t)),\ t \in [1, n]. \tag{10}$$

3.2.2.2 Commonsense knowledge fusion The internal state (q_s), external state (r_s), and intent state (i_s) are passed to a commonsense knowledge fusion module (State Attention). The State Attention contains a multi-head attention fusion module, which helps the network integrate commonsense information. Because the intent state has a more significant guiding role than the internal state and external state in recognizing the emotion of the dialogue, we use the intent state (i_s) as the Query (Q) and Value (V) of the multi-head attention operation as input to the fusion module.

Then, the internal state (i_s) and external state (r_s) are used as the Key (K) of the multi-head attention to adjust the dialogue to commonsense attention at any moment.

$$Q_t = W_Q \cdot i_{s(u_t),t} \tag{11}$$

$$V_t = W_V \cdot i_{s(u_t),t} \tag{12}$$

$$K_t = W_V \cdot (r_{s(u_t),t},\ q_{s(u_t),t}) \tag{13}$$

$$S_t = V_{tseq}^{softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right). \tag{14}$$

3.2.3 Emotion state

The emotional state ultimately determines the speaker's emotion and the emotion of the current utterance. This emotional state is shaped by a combination of internal, external, and intent state, which are weighted and integrated to form a composite emotional state. Additionally, the speaker's previous emotional state can also influence their current emotional state and thus impact the emotion conveyed in their current utterance. The GRU_E formula is shown below:

$$e_t = GRU_E(e_{t-1},\ x_t \oplus S_t),\ t \in [1, n]. \tag{15}$$

3.3 Emotion classifier

Finally, all utterances in the dialogue are classified using a fully connected network.

$$P_t = softmax(W_c e_t + b_c),\ t \in [1, n] \tag{16}$$

$$\hat{y}_t = \underset{k}{argmax}(P_t[k]) \tag{17}$$

The fully connected layer's weight and bias values are denoted as W_c and b_c, respectively. $P_t[k]$ represents the probability of the utterance u_t belonging to class k, while $\hat{y}_t$ indicates the ultimate emotion label assigned to utterance u_t.

4 Experimental setup

4.1 Dataset

We evaluate our model on four datasets: IEMOCAP [47], DailyDialog [48], EmoryNLP [24], and MELD [49]. The statistics of the dataset are shown in Table 1.

IEMOCAP [47]: IEMOCAP consists of 12 hours of two-person conversations between ten unique speakers, of which only the first eight speakers from sessions one through four were used in the training set. Previous work considered six emotions : *neutral*, *happy*, *sad*, *angry*, *excited*, *frustrated*. We follow the same training and validation set split as Ghosal et al.

DailyDialog [48]: DailyDialog is a dataset containing two-way conversations about daily life. It contains seven emotions : *neutral*, *happiness*, *sadness*, *anger*, *surprise*, *disgust*, *fear*. More than 83% of the utterances in this dataset are labeled as neutral.

MELD [49]: MELD is a dataset that contains multi-speaker dialogues from the TV series "Friends". The emotion categories are the same as in DailyDialog.

EmoryNLP [24]: EmoryNLP material also comes from the TV series "Friends", which has three or more speakers in a conversation. The seven emotions are: *neutral*, *mad*, *sad*, *scared*, *powerful*, *peaceful*, *joyful*.

Table 1 Data distribution of IEMOCAP, DailyDialog, EmoryNLP, MELD datasets

Dataset	Number of dialogue			Number of utterance		
	Train	Dev	Test	Train	Dev	Test
IEMOCAP	120		31	5810		1623
DailyDialog	11,118	1000	1000	87,170	8069	7740
EmoryNLP	659	89	79	7551	954	984
MELD	1039	114	280	9989	1109	2610
	Average dialogue length			Average utterance length		
IEMOCAP	48		52	12		13
DailyDialog	8	8	8	12	11	12
EmoryNLP	12	11	13	8	7	8
MELD	10	10	9	8	8	8

4.2 Training setup

To extract context-free features, we fine-tune the RoBERTa model on the set of all utterances and their corresponding sentiment labels in the training data. The fine-tuning process is performed with a batch size of 32 utterances, using the Adam optimizer with a learning rate of 1e−5. To improve the stability of the emotion recognition models trained on the MELD and EmoryNLP datasets, residual connections are used between the first and penultimate layers. The parameter settings for the experiment are as follows: For the MELD dataset, we employ a batch size of 8 and run for 40 training iterations, the dropout is 0.5; For IEMOCAP, we use a batch size of 16 and conduct 60 training iterations, the dropout is 0.25; In the case of the EmoryNLP dataset, our batch size is 16, and we train for 30 iterations, the dropout is 0.5, while for the DailyDialog dataset, we choose a batch size of 8 and proceed with 20 training iterations, the dropout is 0.5. The L2 regularization parameter is 3e−4. The training of the ERC model involves utilizing the Adam optimizer, and a learning rate of 1e−4 is employed.

5 Results and analysis

5.1 Baseline methods

In this section, we compare our proposed model with several benchmark models published in the ERC field in recent years. These models attempt to solve the problem of long-distance context and incorporate commonsense knowledge to enhance dialogue emotion recognition.

CNN [50]: This model stands out as the prevailing choice for employing convolutional neural networks in dialogue emotion recognition.

DialogueRNN [51]: This model divides the speaker in the dialogue into two roles, speaker and listener. Two GRU networks are employed to model the historical discourses of the two roles. Additionally, a third GRU is utilized to aggregate information from the previous two networks through a GRU encoder. This configuration simulates the interaction between the two roles and models global context information.

DialogueGCN [52]: This model uses graph convolutional networks to break the fixed timing relationship in RNN-like modeling. It models the dialogue using the graph structure and uses the mechanism of the graph convolutional network to help target utterances collect information from their neighbor utterances to better judge the sentiment of the target utterance.

HiTrans [53]: This model is a two-layer model based on the Transformer model. The low-level Transformer is utilized to obtain the vector representation of each utterance in the dialogue, while the high-level Transformer is employed to model the context and predict the potential emotions of the utterance. The model includes a speaker relationship prediction task to enhance training intensity.

DIMMN [2]: Designed with a multiview layer, this model fuses the three modalities of text, audio, and video data in a dialogue. Additionally, it considers dynamic interactions among the modalities and proposes the Dynamic Interaction Multiview Memory Network to integrate interaction information among the three modalities for emotion recognition in the conversation.

KET [9]: This model establishes a hierarchical Transformer model using its proposed graph attention mechanism and introduces a commonsense knowledge base to aid in understanding the semantic content of the dialogue.

Sentic GAT [40]: Proposing an attention mechanism for context-aware and emotion-aware graphs embedded in commonsense knowledge, this model maintains the consistency of feelings between commonsense knowledge and the corresponding vocabulary. Additionally, it characterizes the degree of sentiment coherence with the help of emotional intensity.

AutoML-Emo [41]: A Transformer-based model that can be used to capture context and incorporate commonsense knowledge. Based on this, LESM for reducing commonsense knowledge base and GESM based on genetic algorithms are extended to better accommodate datasets of different sizes and domains.

COSMIC [10]: This model is an improved version of DialogueRNN, using bidirectional GRU to encode the speaker and refining the speaker's state, including internal state, external state, contextual state, emotional state, and intentional state. It uses COMET to generate commonsense knowledge for these states.

KI-Net [54]: This model combines sentiment lexicon and commonsense knowledge to enhance semantic representation. A self-matching module is used for internal utterance-knowledge interaction.

TODKAT [39]: This model captures the meaning of discourse in different topic-driven contexts and incorporates external commonsense knowledge to model the interlocutor's intentions and behaviors. It transforms text classification into an encoder-decoder form of generative model and further improves the accuracy of dialogue emotion recognition.

5.2 Comparison with the benchmark model

Following previous works, for IEMOCAP, EmoryNLP, and MELD, we select weighted F1 score as the evaluating metric. For DailyDialog, we select the micro F1 score excluding those utterances labeled with neutral.

Table 2 presents the experimental results of our proposed method compared to benchmark baselines. Our model achieves state-of-the-art results on the IEMOCAP and DailyDialog datasets and competitive results on the MELD and EmoryNLP datasets compared to other models. To demonstrate the importance of selecting commonsense knowledge, we have included additional models that use commonsense knowledge for comparison.

In comparison to the sequential model DialogueRNN, the graph structure model DialogueGCN performs better. This suggests that modeling context with a graph network is superior to that of the sequential model (RNN). After that, HiTrans started to use the two-layer Transformer structure to model the context. Although these models use different approaches to construct the context inside the dialogue, they all lack the injection of commonsense knowledge outside the dialogue and thus perform poorly compared to our model.

Comparing with DIMMN, their model extracts more information from the feature interactions among the three modalities that are useful for analyzing conversations, but they also lack help from context as well as external knowledge. It can be seen that for ERC task, the global contextual information of textual modality and the injection of knowledge have some advantages over multimodal modeling.

When compared to the KET model, which also incorporates external commonsense knowledge, our proposed model performs significantly better. This is because we classify commonsense knowledge into three states and assign higher weight to the state that can play a more significant role in guiding dialogue emotion recognition. This demonstrates that the strategic use of commonsense knowledge can enhance the optimization effect of commonsense knowledge on dialogue emotion recognition.

Table 2 ERC performance (F1 score) of different approaches on IEMOCAP, DailyDialog, EmoryNLP and MELD datasets

Methods	IEMOCAP	DialyDialog	EmoryNLP	MELD
CNN	52.04	50.32	32.59	55.02
DialogueRNN	62.57	55.95	31.7	57.03
DialogueGCN	64.18	–	–	58.1
HiTrans	64.5	–	36.75	61.94
DIMMN	64.1	–	–	58.6
KET#	59.56	53.37	34.39	58.18
Sentic GAT#	–	53.34	35.27	56.86
AutoML-Emo#	–	54.82	35.77	58.66
COSMIC#	65.28	58.48	38.11	65.21
KI-Net#	–	57.3	–	63.24
TODKAT#	62.75	58.47	**38.69**	**65.47**
Ours	**65.74**	**59.02**	38.65	65.28

Among them, KET, Sentic GAT, AutoML-Emo, COSMIC, KI-Net and TODKAT are the application of external knowledge. Symbol '–' indicates that no results are given in the corresponding paper. Symbol '#' indicates that these methods use external commonsense knowledge

Bold values indicate highest score compared to other results

Comparing with the experimental results of Sentic GAT and AutoML-Emo, our experimental results have a clear advantage. Although all three perform context-specific selection of commonsense knowledge, the local–global context encoder of our model takes into account the interaction of distant utterances, neighboring utterances, and intra-utterance information. In addition, these two models consider the selection of commonsense knowledge more at the sentiment word level, while our model selects the extracted commonsense knowledge from the perspective of modeling the speaker's mental state, which is more accurate in predicting utterance emotion.

In contrast to COSMIC, which has limited improvement on RoBERTa DialogueRNN on the context-rich IEMOCAP dataset, our model more effectively combines intra-sentence, short-range, and long-range contextual semantic information, demonstrating our model's superiority in context modeling. As shown in Fig. 1, the interaction between long-distance utterances plays a crucial role in MELD. Moreover, our proposed model has improved on the EmoryNLP, DailyDialog, and MELD datasets, indicating that the weighted selection of commonsense knowledge is effective.

In comparison to KI-Net, we have achieved better results in two datasets. Despite their innovations in the interaction of information inside a conversation with external knowledge, they lack the integration of local context with global context when dealing with information inside a conversation, which results in not providing enough information for predicting emotions in a conversation.

Compared to the best performance model TODKAT, we outperform it on the IEMOCAP and DailyDialog datasets, but fall short on the EmoryNLP and MELD datasets. TODKAT first designed an additional layer dedicated to topic detection, and then combined topic-driven information with commonsense knowledge based on dialogue context information. Experimental results show that combining commonsense knowledge with topic-driven information is more effective than our method in multi-person dialogues, but not as effective in two-person dialogues. Compared to our model, the topic-driven approach proposed by TODKAT provides a clearer target for complex semantic information in the datasets MELD and EmoryNLP with multiple people participating in the dialogue, which is conducive to modeling semantic dependencies. However, in the two-person dialogue datasets IEMOCAP and DailyDialog, the speaker's psychological state has a greater impact on emotion than the topic, and our model's weighted selection of different commonsense states can better describe the speaker's psychological state. In the future, topic-driven approaches can be incorporated into our model to improve its ability to model semantic dependencies.

5.3 Ablation experiment

We need to evaluate the impact of different modules on our model. We removed a relation from our proposed model to test the effect of the removed module on the model's performance.

As shown in Table 3, we have conducted three relevant ablation experiments on the model proposed in this paper, respectively. By removing the global context encoder from the model and letting the local context information extracted by RoBERTa serve as context state, we can observe that the performance of the model on all the four datasets is attenuated to different degrees. This suggests that local context encoder can effectively model the relative positions of the utterances in a conversation. After that, we also tried to remove the local–global context coder from the model, so that the vectors obtained from local context coder and global context coder can only be used to extract commonsense knowledge features and as context state, respectively. The performance of the model obtained in this way took a very serious dip, showing that the loss of local–global context coder deprives the model of its ability to integrate information from inside and outside the conversation. Finally, we also tried removing the commonsense knowledge fusion module after extracting the commonsense knowledge and only introducing three commonsense states. However, we found that the model's performance still degraded on all four datasets, which suggests that choosing the appropriate general knowledge feature extraction is crucial for improving the accuracy of dialogue emotion recognition.

In Fig. 3, we show the effect of increasing the number of speakers in a conversation on the dataset MELD for both our method and the method that removes the commonsense knowledge fusion module (which simply injects commonsense knowledge, similar to COSMIC). As the number of speakers increases, both our method and the method that mechanically injects commonsense knowledge initially show improved performance. Our approach achieves competitive performance compared to without commonsense knowledge fusion approach. This suggests that our method can have outstanding performance when dealing with a small number of speakers. However, when the number of speakers continues to increase, our method shows the same performance degradation as the unselected method. By the time the number of speakers comes to 8, our method is already below simple common sense injection. This suggests that when a conversation involves a large number of speakers, it becomes difficult for our method to be effective.

Table 3 Performance (F1-score) of our model and variants with some modules removed on four datasets

Method	IEMOCAP	DailyDialog	EmoryNLP	MELD
Ours	**65.74**	**59.02**	**38.65**	**65.28**
w/o Global context encoder	65.42	58.77	38.4	65.15
w/o Local–global context encoder	64.8	57.95	37.93	64.32
w/o Commonsense knowledge fusion	65.33	58.52	38.53	65.03

Bold values indicate highest score compared to other results

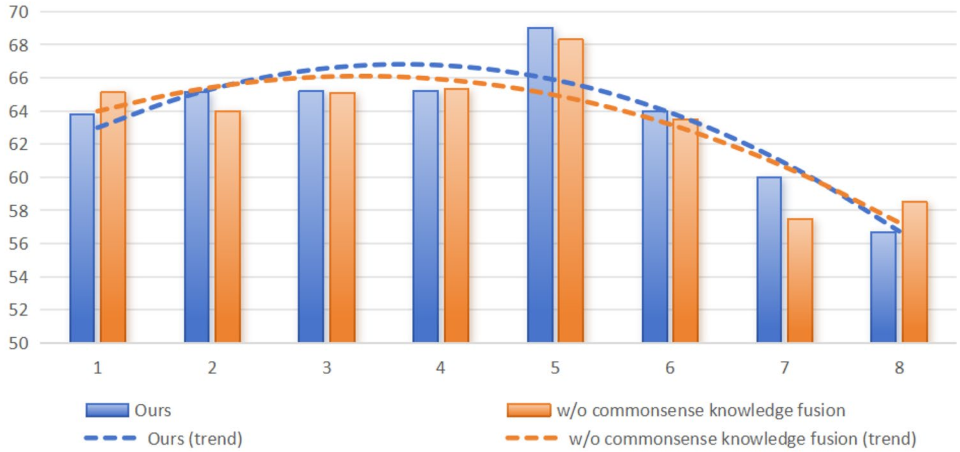

Fig. 3 F1 scores of conversations with different number of speakers on MELD achieved by our method and without commonsense knowledge fusion. X-axis denotes F1 score; Y-axis denotes the number of speakers in a dialogue. The dotted line denotes the trend line of second-order polynomial

5.4 Impact of different kinds of commonsense knowledge states

To determine which commonsense knowledge state plays a more important role in guiding dialogue emotion recognition, we respectively used the three emotional states as the parameter Q and V inputs of multi-head attention, and evaluated their effect on the four datasets.

Table 4 shows that on the IEMOCAP and DailyDialog datasets, using the intent state as the multi-head attention Q, V input slightly reduces the model's performance compared to the external state, while outperforming both the external and internal states on the MELD and EmoryNLP datasets. We hypothesize that in two-person dialogue datasets such as IEMOCAP and DailyDialog, other modalities related to the external state, such as external expressions, speech patterns, acoustic features, visual expressions, gestures, and postures, are more easily perceived by the other speaker. In multi-person dialogue datasets such as MELD and EmoryNLP, the internal and external states of the speaker may be less clear, while the intent state may be more closely related to the speaker's psychological state and have a stronger influence on the dialogue emotion. Therefore, using the intent state as the multi-head attention Q, V input better models the speaker's psychological state, leading to a more accurate prediction of dialogue emotion.

Table 4 Select the q_s, r_s, and i_s commonsense knowledge states as the performance(F1-score) of the Q, V inputs of the multi-head attention on the four datasets

Q,V	IEMOCAP	DailyDialog	EmoryNLP	MELD
q_s	65.69	58.17	37.96	64.29
r_s	**65.82**	**59.47**	38.22	64.55
i_s	65.74	59.02	**38.65**	**65.28**

Bold values indicate highest score compared to other results

5.5 Case study

To demonstrate the effectiveness of our model in combining contextual semantic dependencies in the conversation and external commonsense knowledge outside the conversation, we present a typical case study from the MELD dataset, as shown in Fig. 4. Past models have tended to make incorrect predictions for such cases with several emotion-shifts between utterances, because they ignored the help of distant context as well as commonsense knowledge.

In this conversation, after some columns, there are some emotion-shifts(neutral, anger) between utterances, and we accurately predict the emotion of the target utterance by fusing commonsense states from long-distance, short-distance context, and intra-utterance. Utterances #9, #18, #20, and #22, all spoken by Rachel, show a clear change in her mental state—from initial disappointment, feeling abandoned by Joey, to a final gradual transformation into an angry emotional state.

5.6 Error analysis

Figures 5, 6, 7, and 8 depict the confusion matrices generated by our model on the four datasets: EmoryNLP, MELD, DailyDialog, and IEMOCAP. The emotional labels on the x-axis are manually labeled, while the emotional

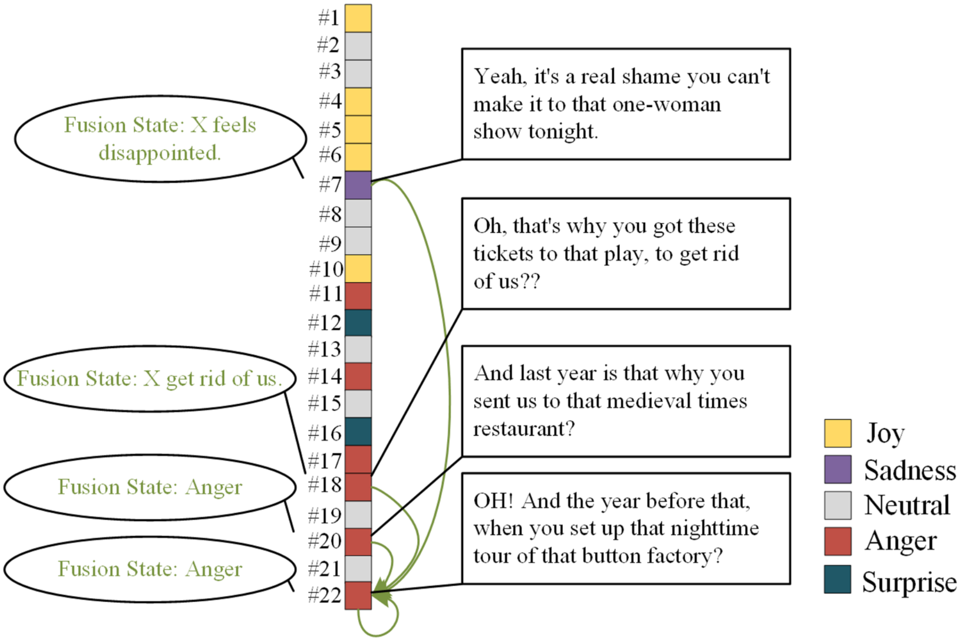
Fusion State: X feels disappointed.
Fusion State: X get rid of us.
Fusion State: Anger
Fusion State: Anger
#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14
#15
#16
#17
#18
#19
#20
#21
#22
Yeah, it's a real shame you can't make it to that one-woman show tonight.
Oh, that's why you got these tickets to that play, to get rid of us??
And last year is that why you sent us to that medieval times restaurant?
OH! And the year before that, when you set up that nighttime tour of that button factory?
Joy
Sadness
Neutral
Anger
Surprise

Fig. 4 Case study from the dataset MELD. Utterances on the left and selected commonsense knowledge on the right

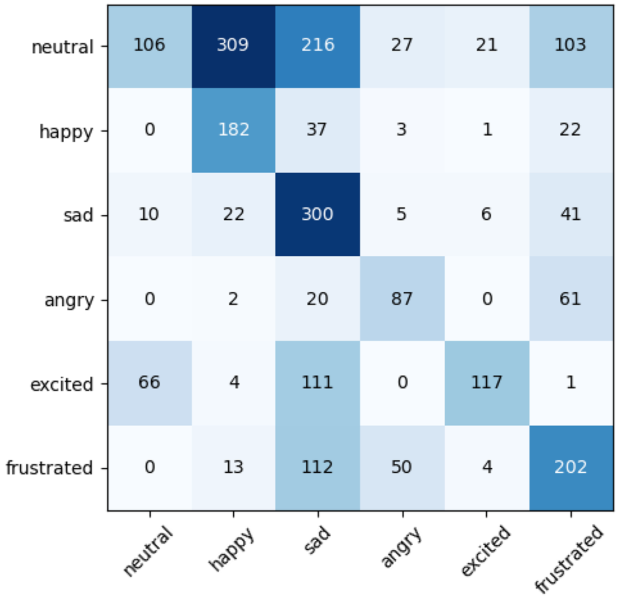

Fig. 5 Confusion matrix of IEMOCAP dataset

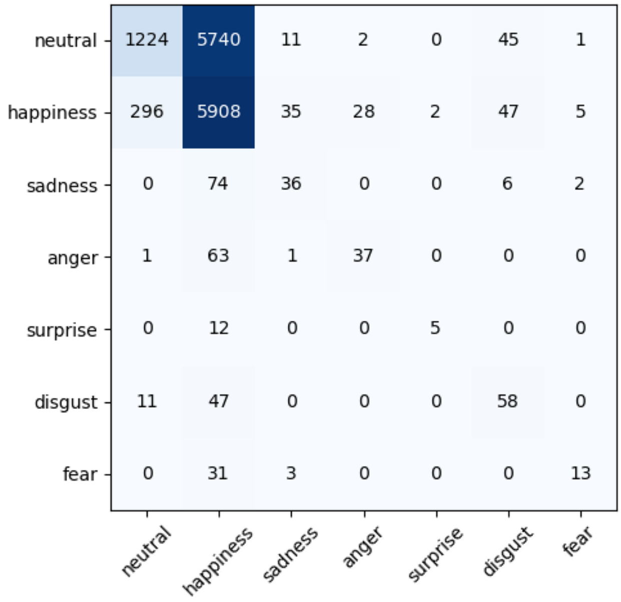

Fig. 7 Confusion matrix of DailyDialog dataset

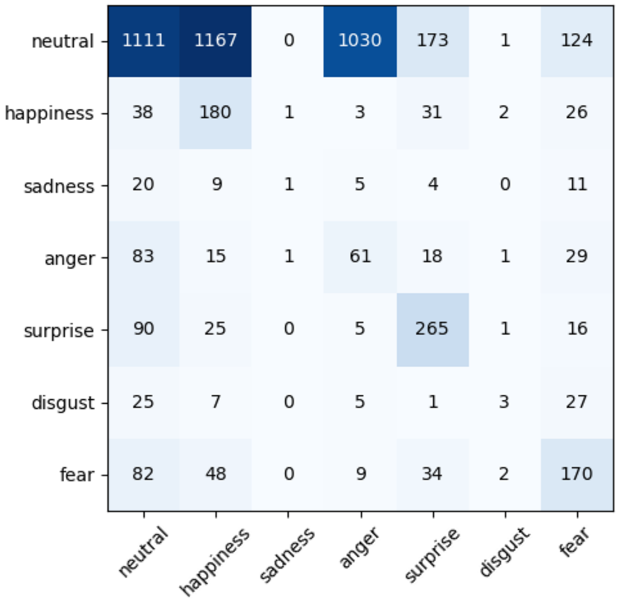

Fig. 6 Confusion matrix of MELD dataset

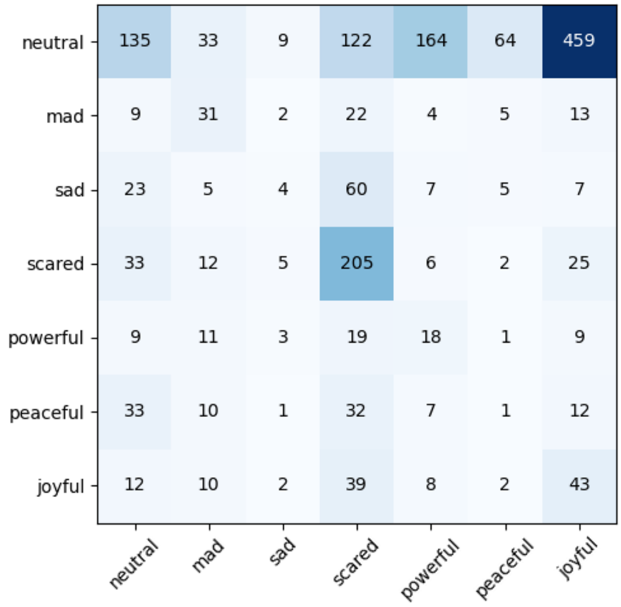

Fig. 8 Confusion matrix of EmoryNLP dataset

labels on the y-axis represent the predictions made by our model. Except for the diagonal line starting from the upper-left corner, the intensity of the color represents the severity of emotion prediction errors at the corresponding position. The heat maps reveal that samples of positive emotions, such as *happy*, *happiness*, and *joy*, as well as negative samples, such as *anger*, are frequently predicted as *neutral*. These observations indicate that distinguishing between *neutral* and some less emotional samples in emotion recognition remains challenging for our approach.

Furthermore, we observed errors in predicting specific samples. For example, the utterance "Oh, get a room." is incorrectly categorized as *joy* when its correct emotion label should be *disgust*. This misclassification occurs because our model relies solely on contextual and commonsense knowledge, while the MELD dataset is multimodal. The video feature of the utterance shows Phoebe feeling disgusted because Chandler and Monica are kissing in front of her; it is context-independent, whereas this type of utterance requires a combination of video and audio features to correctly predict emotion.

6 Conclusion

In this paper, we propose a local–global context commonsense fusion model for dialogue emotion recognition, enabling more effective utilization of long-distance context information and commonsense knowledge. We obtain intra-utterance, local contextual, and global contextual information through the local context encoder and the global context encoder. The commonsense fusion module uses a weighting mechanism to select the most relevant external commonsense. The local–global context encoder effectively models the complex semantic dependencies among the current utterance, local context, and global context. The commonsense knowledge fusion module not only enriches the semantic information but also reduces noise information by describing the speaker's psychological state. Our method achieves competitive performance compared to benchmark models on four widely used public datasets. By optimizing the modeling of semantic dependencies and carefully incorporating commonsense knowledge, our approach provides a promising direction for improving ERC model performance.

In future work, we plan to introduce multi-modal information and optimize its combination to further enrich the semantic information of the dialogue. Additionally, describing psychological states through the weight selection of commonsense is somewhat imprecise, and alternative methods may replace multihead-attention. In the context of multi-person dialogue, there are still many opportunities to optimize the modeling of semantic dependencies, such as causal reasoning or the topic-driven approach used in the comparison model (TODKAT).

Acknowledgements The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Science and Technology Major Project of Guangdong Province (STKJ2021005, STKJ202209002, STKJ2023076), Natural Science Foundation of Guangdong Province (2019A1515010943).

Authors' contributions W.Y.: Original Draft. C.L.: Review & Editing. X.H.: Review & Editing. W.Z.: Validation. E.C.: Review & Editing. D.J.: Supervision.

Funding The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Natural Science Foundation of Guangdong Province (2019A1515010943), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence)(2019KZDZX1030), 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D), Science and Technology Major Project of Guangdong Province(STKJ2021005, STKJ202209002), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology (2020B1212060078).

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Kratzwald B, Ilic S, Kraus M, Feuerriegel S, Prendinger H (2018) Decision support with text-based emotion recognition: deep learning for affective computing. arXiv preprint arXiv:1803.06397
2. Wen J, Jiang D, Tu G, Liu C, Cambria E (2023) Dynamic interactive multiview memory network for emotion recognition in conversation. Inf Fusion 91:123–133
3. Cambria E, Wang H, White B (2014) Guest editorial: big social data analysis. Knowl-Based Syst 69:1–2
4. Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167
5. Saberi B, Saad S (2017) Sentiment analysis or opinion mining: a review. Int J Adv Sci Eng Inf Technol 7(5):1660–1666
6. Baecker AN, Geiskkovitch DY, González AL, Young JE (2020) Emotional support domestic robots for healthy older adults: conversational prototypes to help with loneliness. In: Companion of the 2020 ACM/IEEE international conference on human–robot interaction, pp 122–124
7. Abdollahi H, Mahoor MH, Zandie R, Sewierski J, Qualls SH (2022) Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. IEEE Trans Affect Comput 14(3):2020–2032. https://doi.org/10.1109/TAFFC.2022.3143803
8. Darling K (2016) Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In: Law Robot, Froomkin Calo, Kerr (eds) Edward Elgar.
9. Zhong P, Wang D, Miao C (2019) Knowledge-enriched transformer for emotion detection in textual conversations. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 165–176
10. Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S (2020) COSMIC: COmmonSense knowledge for eMotion identification in conversations. In: Findings of the association for computational linguistics: EMNLP 2020, pp 2470–2481
11. Li J, Lin Z, Fu P, Wang W (2021) Past, present, and future: conversational emotion recognition through structural modeling of psychological knowledge. In: Findings of the association for computational linguistics: EMNLP 2021, pp 1204–1214
12. Hu J, Liu Y, Zhao J, Jin Q (2021) 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), pp 5666–5675

13. Wen Z, Wang R, Luo X, Wang Q, Liang B, Du J, Yu X, Gui L, Xu R (2023) Multi-perspective contrastive learning framework guided by sememe knowledge and label information for sarcasm detection. Int J Mach Learn Cybern 14:4119–4134
14. Wang R, Bao J, Mi F, Chen Y, Wang H, Wang Y, Li Y, Shang L, Wong K-F, Xu R (2023) Retrieval-free knowledge injection through multi-document traversal for dialogue models. In: Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 6608–6619
15. Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y (2019) COMET: commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 4762–4779
16. Xiao G, Tu G, Zheng L, Zhou T, Li X, Ahmed SH, Jiang D (2020) Multimodality sentiment analysis in social internet of things based on hierarchical attentions and CSAT-TCN with MBM network. IEEE Internet Things J 8(16):12748–12757
17. Jiang D, Liu H, Wei R, Tu G (2023) CSAT-FTCN: a fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition. Cogn Comput 15:1082–1091
18. Tu G, Wen J, Liu H, Chen S, Zheng L, Jiang D (2022) Exploration meets exploitation: multitask learning for emotion recognition based on discrete and dimensional models. Knowl-Based Syst 235:107598
19. Khan W, Daud A, Nasir JA, Amjad T (2016) A survey on the state-of-the-art machine learning models in the context of NLP. Kuwait J Sci 43(4):95–113
20. Tu G, Liang B, Jiang D, Xu R (2022) Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations. IEEE Trans Affect Comput 14:1803–1816
21. Chen R, Wang J, Yu L-C, Zhang X (2023) Decoupled variational autoencoder with interactive attention for affective text generation. Eng Appl Artif Intell 123:106447
22. Sheng D, Wang D, Shen Y, Zheng H, Liu H (2020) Summarize before aggregate: a global-to-local heterogeneous graph inference network for conversational emotion recognition. In: Proceedings of the 28th international conference on computational linguistics, pp 4153–4163
23. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P (2017) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long Papers), pp 873–883
24. Zahiri SM, Choi JD (2018) Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In: Workshops at the thirty-second AAAI conference on artificial intelligence
25. Ishiwatari T, Yasuda Y, Miyazaki T, Goto J (2020) Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 7360–7370
26. Zhang D, Wu L, Sun C, Li S, Zhu Q, Zhou G (2019) Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In: IJCAI, pp 5415–5421
27. Shen W, Wu S, Yang Y, Quan X (2021) Directed acyclic graph network for conversational emotion recognition. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 1551–1560
28. Hu D, Wei L, Huai X (2021) 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), pp 7042–7052
29. Lee J, Lee W (2022) CoMPM: context modeling with speaker's pre-trained memorytracking for emotion recognition in conversation. In: Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: human language technologies, pp 5669–5679
30. Wang Y, Zhang J, Ma J, Wang S, Xiao J (2020) Contextualized emotion recognition in conversation as sequence tagging. In: Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue, pp 186–195
31. Chen R, Wang J, Yu L-C, Zhang X (2023) Learning to memorize entailment and discourse relations for persona-consistent dialogues. arXiv preprint arXiv:2301.04871
32. Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 31
33. Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith NA, Choi Y (2019) Atomic: an Atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 3027–3035
34. Cambria E, Liu Q, Decherchi S, Xing F, Kwok K (2022) Senticnet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: Proc LREC 2022, pp 3829–3839
35. Cai H, Shen X, Xu Q, Shen W, Wang X, Ge W, Zheng X, Xue X (2023) Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. arXiv preprint arXiv:2306.04657
36. Liu Y, Wan Y, He L, Peng H, Philip SY (2021) Kg-bart: knowledge graph-augmented bart for generative commonsense reasoning. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 6418–6425
37. Zhang X, Bosselut A, Yasunaga M, Ren H, Liang P, Manning CD, Leskovec J (2022) Greaselm: graph reasoning enhanced language models for question answering. arXiv preprint arXiv:2201.08860
38. Song R, He S, Gao S, Cai L, Liu K, Yu Z, Zhao J (2023) Multilingual knowledge graph completion from pretrained language models with knowledge constraints. In: Findings of the association for computational linguistics: ACL 2023, pp 7709–7721
39. Zhu L, Pergola G, Gui L, Zhou D, He Y (2021) Topic-driven and knowledge-aware transformer for dialogue emotion detection. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 1571–1582
40. Tu G, Wen J, Liu C, Jiang D, Cambria E (2022) Context- and sentiment-aware networks for emotion recognition in conversation. IEEE Trans Artif Intell 3(5):699–708
41. Jiang D, Wei R, Wen J, Tu G, Cambria E (2023) AutoML-Emo: automatic knowledge selection using congruent effect for emotion identification in conversations. IEEE Trans Affect Comput 14:1845–1856
42. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
43. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186
44. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

45. Liu Y, Lapata M (2019) Text summarization with pretrained encoders. 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), pp 3730–3740
46. Radford A, Narasimhan K, Salimans T, Sutskever I, et al (2018) Improving language understanding by generative pre-training
47. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
48. Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017) DailyDialog: a manually labelled multi-turn dialogue dataset. In: Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers), pp 986–995
49. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: ACL, pp 527–536
50. Chen Y (2015) Convolutional neural network for sentence classification. Master's thesis, University of Waterloo
51. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) Dialoguernn: an attentive RNN for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence, pp 6818–6825
52. Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A (2019) DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 154–164
53. Li J, Ji D, Li F, Zhang M, Liu Y (2020) Hitrans: a transformer-based context-and speaker-sensitive model for emotion detection in conversations. In: Proceedings of the 28th international conference on computational linguistics, pp 4190–4200
54. Xie Y, Yang K, Sun C-J, Liu B, Ji Z (2021) Knowledge-interactive network with sentiment polarity intensity-aware multi-task learning for emotion recognition in conversations. In: Findings of the association for computational linguistics: EMNLP 2021, pp 2879–2889

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.