DetectiveNN: Imitating Human Emotional Reasoning with a Recall-Detect-Predict Framework for Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Recognizing emotions in conversations involves an internal cognitive process that interprets emotional cues by using a collection of past emotional experiences. However, many existing methods struggle to decipher emotional cues in dialogues due to their models' lack of capacity for cognitive reasoning. In this work, we introduce an innovative Detective Network (DetectiveNN), a novel model that is grounded in the cognitive theory of emotion and utilizes a "recall-detect-predict" framework to imitate human emotional reasoning. This process begins by 'recalling' past interactions of a specific speaker to collect emotional cues. It then 'detects' relevant emotional patterns by interpreting these cues in the context of the ongoing conversation. Finally, it 'predicts' the speaker's emotional state in the next moment. Tested on three benchmark datasets, our approach significantly outperforms existing methods. This highlights the advantages of incorporating cognitive factors into deep learning, enhancing task efficiency and prediction accuracy[1].

1 Introduction

In recent years, recognizing emotions in dialogues has gained increasing attention in the field of natural language processing (NLP). This surge in attention is driven by its vast potential for application in areas like human-computer interaction and empathetic dialogue systems (Ma et al., 2020; Concannon and Tomalin, 2023)

In the realm of conversational emotion recognition (ERC), interpreting emotional cues embedded in conversational context is crucial (Mittal et al., 2020; Gomathy, 2021). Conversations are filled with emotional cues that act as triggers for the emotions expressed in a current utterance (Oberländer et al., 2020; Hu et al., 2021). ERC seeks to detect and interpret these emotional clues within the flow of conversation, aiming for a nuanced understanding of the emotional context. Traditional ERC approaches typically adopt a 'recall-then-predict' strategy (Mitra et al., 2023), modeling both speaker-level and dialogue-level contexts to predict emotional states in conversations. DialogueGCN (Ghosal et al., 2019) models interactions between speakers using graph networks to capture emotional cues throughout the conversation. DialogXL (Shen et al., 2021) introduces a dialog-aware self-attention mechanism within a transformer structure to capture emotional cues, including intra- and inter-speaker dependencies. C-LSTM (Zhou et al., 2015) leverages a LSTM-based approach to encode the global context, whereas DialogueRNN (Majumder et al., 2019) employs GRUs to track both speaker state and global state for each conversation. COSMIC (Ghosal et al., 2020) leverages external commonsense knowledge to enhance the model's ability to detect rich emotional cues. Additionally, DialogueCRN (Hu et al., 2021) employs a multi-turn reasoning module that extracts and integrates emotional clues from the dialogue. Although termed 'reasoning,' this process fundamentally involves retrieving and combining contextual clues at each turn before classifying the emotion. Existing models face challenges in accurately analyzing and decoding emotional cues, primarily due to the absence of a cognitive reasoning phase.

Emotion recognition can be understood as the process of deciphering emotional cues to comprehend the cognitive context, aligning with the Cognitive Theory of Constructed Emotion (Russell, 2003, 2009; Barrett and Russell, 2014). This theory suggests that emotions are formed from an individual's cognitive context, shaped by their thoughts, memories, and social interactions (Barrett, 2014). Inspired by this theory, we approach the ERC tasks as an internal cognitive process that deciphers each participant's emotional cues based on their past emotional experiences in a dialogue. This process

[1] our code can be found here

involves identifying and organizing emotional cues, synthesizing them into a coherent emotional narrative, and subsequently examining this narrative throughout the conversational context to validate the cues. We propose a novel Detective Network (DetectiveNN) with a 'recall-detect-predict' strategy for enhanced ERC accuracy. The DetectiveNN model features a detection phase that deciphers emotional cues throughout the conversation context, connecting these cues to decode the evolution of a speaker's emotional responses. This phase reveals patterns in a speaker's emotional flows, akin to a detective piecing together clues to map an individual's emotional states.

DetectiveNN begins with a recall phase, where we utilize a sequence-based model to retrieve contextual information from the personal emotional experiences and interactions of each speaker. This approach is inspired by the pioneering work of Hu (Hu et al., 2021) and Yang (Yang and Shen, 2021), who demonstrated the efficacy of sequence-based models in learning diverse contextual information.

In the detection phase, we employ a transformer-like architecture to iteratively analyze and decode emotional cues drawn from the extensive emotional experiences of a specific speaker. This phase is divided into two key operations: an examination process and a conscious detection process. The examination process utilizes transformer encoders to both identify and clarify the logical connections between emotional cues, effectively merging these cues into a coherent emotional narrative. It achieves a deep understanding of the speaker's emotional context by integrating cues through a set of encoder blocks. The conscious detection process employs a cross-attention mechanism, probing the speaker's constructed emotional narrative and integrating the dynamic interplay between emotional cues and the speaker's historical interactions. This method uncovers patterns that decode the speaker's emotional journey, offering insights into the evolution of emotional states over time.

Following the insights gained from the detection phase, an emotion classifier predicts the emotion label of each utterance. By incorporating the 'recall-detect-predict' framework, DetectiveNN effectively mirrors the cognitive reasoning process humans use to understand emotional states. We hypothesize that integrating cognitive reasoning into deep learning models significantly enhances their capability to analyze and interpret emotions in each dialogue segment.

To assess the efficacy of our proposed model, extensive experiments were conducted on three widely accepted benchmark datasets: IEMOCAP, EmoryNLP and Dailydialog. The experimental results demonstrate that our model significantly outperforms existing methods, primarily attributed to the application of a cognitive approach in deciphering emotional cues.

The primary contributions of our research are as follows:

- We introduce an innovative Detective Network (DetectiveNN) designed within a 'recall-detect-predict' framework, drawing on principles of cognitive theory of constructed emotion.
- We design a transformer architecture to perform the detection process. This architecture plays a key role in interpreting emotional cues in conversations, enhancing the accuracy and nuances of recognizing different emotions in dialogues.
- We conduct extensive experiments on three benchmark datasets. The results consistently demonstrate the effectiveness and superiority of the proposed model (see Figure 1).

2 Related Work

The ERC field has significantly advanced, with an emphasis on extracting and integrating emotional clues from conversations. This progress can be grouped into three major methodologies: Sequence-based models, Pre-trained Language Model-based Models and Graph-based Models.

Sequence-based models: DialogueRNN (Majumder et al., 2019) utilizes three gated recurrent units (GRUs) to track emotional states in conversations by integrating speaker identity, context, and emotions from neighboring utterances to maintain individual speaker states. DialogueCRN (Hu et al., 2021) integrates cognitive theories of emotion, featuring a reasoning module for iterative extraction and integration of emotional clues, combining intuitive retrieving (attention mechanisms) and conscious reasoning using Long Short-Term Memory (LSTM) network. BC-LSTM (Poria et al., 2018) captures both the left and right context of each utterance using bidirectional LSTMs, understanding the influence of preceding and following utterances. CMN (Hazarika et al., 2018b) utilizes

a multimodal approach to model past utterances of each speaker into memories with GRUs. EmotionIC (Yingjian et al., 2023) employs a three-part framework: Identity Masked Multi-Head Attention (IMMHA) to grasp the overall context, Dialogue-based Gated Recurrent Unit (DiaGRU) to capture immediate conversational nuances, and Skip-chain Conditional Random Field (SkipCRF) to trace the progression of emotions, thus integrating attention with recurrence for comprehensive emotion detection in dialogue. COSMIC (Ghosal et al., 2020) integrates commonsense knowledge with GRUs to model various states of conversation, including the internal, external, intent, and emotional states of speakers.

Pre-trained Language Model-based Models: DialogXL (Shen et al., 2021) employs XLNet (Yang et al., 2019) to process longer conversational histories and incorporates dialogue-level self-attention to manage multi-party conversation dynamics effectively. Emoberta (Kim and Vossen, 2021) uses RoBERTa (Liu et al., 2019) to predict the emotion of a current speaker by learning speaker-level and dialogue-level context.

Graph-based Models: DialogueGCN (Ghosal et al., 2019) utilizes a graph convolutional neural network to model conversational context. It represents utterances as nodes in a graph, capturing dependencies between utterances as edges for a better understanding of conversational dynamics. Zhang et al. (2019) employs a graph convolutional neural network to capture the context-sensitive dependence between utterances in the same conversation and the speaker-sensitive dependence between an utterance and its speaker node. Lian et al. (2020) utilizes a graph neural network with an attention mechanism to model utterance-level and speaker-level context.

3 Methodology

3.1 Problem Definition

We define a conversation consisting of a total number of N utterances. Each utterance in the conversation is associated with a specific speaker. There are S distinct speakers in the conversation. For each speaker, we have a subset of utterances corresponding to this speaker.

The objective of the ERC task is to predict the emotion label for each utterance from the set of emotional labels $\{y_1, y_2, \ldots, y_P\}$ where P is the number of emotional labels.

3.2 Recall Phase

In the realm of ERC, the intra-context is crucial for understanding the emotional journey and thematic progressions of each speaker within their dialogue contributions.

We first utilize a bi-directional GRU network to gather emotional cues and information from utterances generated by speaker s. Each utterance is represented by a feature embedding $x_i \in \mathbb{R}^{du}$, where du is the embedding dimension of each utterance. The sequence of these embeddings is processed by the GRU, with $i = \Phi(k, s)$ mapping the k-th step in the GRU to the corresponding utterance index for the speaker s.

$$c_i^{\text{intra}}, h_{s,k}^{\text{intra}} = GRU^{\text{intra}}(x_i, h_{s,k-1}^{\text{intra}}) \quad (1)$$

where $c_i^{\text{intra}} \in \mathbb{R}^{2du}$ represents an intra-context embedding, and $h_{s,k}^{\text{intra}}$ is the hidden state of the GRU after processing the k-th step for the speaker s.

We sequentially process each c_i^{intra} and compile them into a matrix $C_s^{\text{intra}} \in \mathbb{R}^{N_s \times 2du}$. N_s is the total number of utterances spoken by the speaker s. This matrix builds up as we go through the steps, eventually leading to the final state.

To obtain the global context embedding c_j^{global} representing all interactions between interlocutors, we employ another bi-directional GRU model to capture sequential dependencies between adjacent utterances of interlocutors. The context representation can be computed as:

$$c_j^{\text{global}}, h_j^{\text{global}} = GRU^{\text{global}}(x_j, h_{j-1}^{\text{global}}) \quad (2)$$

where j is an utterance index from the conversation. Similarly we concatenate c_j^{global} to form the matrix $C^{\text{global}} \in \mathbb{R}^{L \times 2du}$. h_j^{global} is the j-th global hidden state of the GRU.

3.3 Detection Phase

The detection phase offers a systematic method for analyzing the underlying emotional dynamics of the speaker s. Initially, it identifies and organizes emotional cues in a logical order. It then synthesizes those cues to form a coherent emotional narrative. Subsequently, the detection phase examines the emotional narrative against the context of the entire conversation, aiming to validate those initial emotional cues. Throughout this analysis, it uncovers patterns in the emotional flows of the speaker s, akin to a detective connecting dots to reveal a broad map of an individual's emotional states.

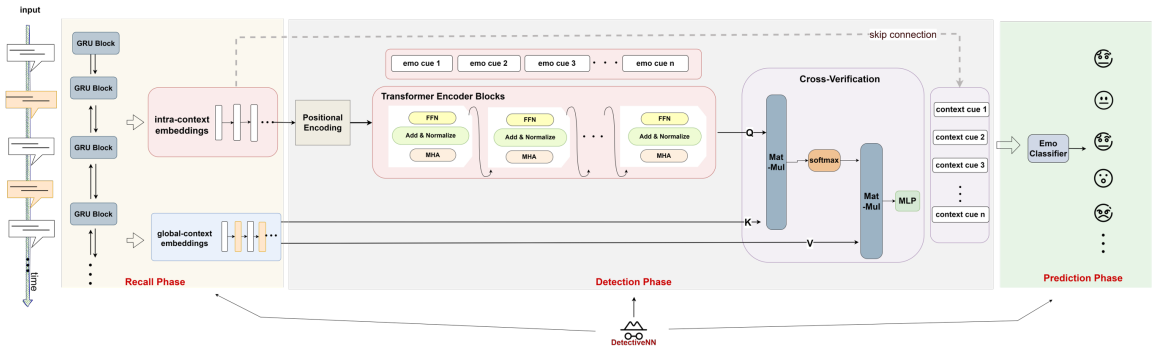

Figure 1: The architecture of the proposed model DetectiveNN

Positional Encoding: We first apply positional encoding, denoted as PE, to inject ordering information to the intra-context matrix C_s^{intra}. This ensures that the DetectiveNN not only processes the inherent emotional cues at each step but also understands its sequential context within the entire process.

We adopt transformer encoder blocks with each block consisting of a Multi-Head Attention layer and a Feed-Forward Network layer to identify and integrate emotional cues from the intra-context.

Multi-Head Self-Attention (MHA) Layer: Our architecture includes an MHA layer with four heads to process the intra-context embedding C_s^{intra}. This layer functions as a detective examining the context of speaker s with each head focusing on different aspects of the emotional content in the speaker's utterances. MHA ensures a thorough, multi-faceted analysis by capturing emotional cues from the intra-context.

Feed-Forward Network (FFN) Layer: Building on the raw emotional cues identified by the MHA layer, the FFN analyzes how those cues interact and connect. Similar to a detective piecing together different clues in a story, the FFN layer builds a comprehensive emotional narrative of speaker s.

Therefore we obtain $\tilde{C}_s^{\text{intra}}$ as the representation of the emotional narrative. It can be expressed as follows:

$$\tilde{C}_s^{\text{intra}} = \text{FFN}\Big(\text{MHA}(C_s^{\text{intra}} + \text{PE}(C_s^{\text{intra}}))\Big) \quad (3)$$

where $\tilde{C}_s^{\text{intra}} \in \mathbb{R}^{N_s \times dc}$. dc is the embedding dimension of the emotional narrative.

Cross-Verification Layer: The DetectiveNN then connects the dots by examining the derived emotional narrative against a broad conversational context. Through careful evaluation, the model identifies patterns in the emotional flows of speaker s. We employ a cross-attention mechanism to mirror this progress. The emotional narrative $\tilde{C^{\text{intra}}}_s$ is treated as a query Q to retrieve additional contextual information from past interactions between the speakers. We set the global context matrix C^{global} as both Key K and Value V.

$$\hat{C}_s^{\text{intra}} = \text{Softmax}\left(\frac{\tilde{C}_s^{\text{intra}} C^{\text{global}^T}}{\sqrt{dc}}\right) C^{\text{global}} \quad (4)$$

where $\hat{C}_s^{\text{intra}} \in \mathbb{R}^{N_s \times dc}$ represents emotional patterns captured through cross verification.

3.4 Emotion Prediction

After retrieving and reasoning emotional clues, the detective is to piece together the puzzle in a way to assess the current emotional state of speaker s.

The emotion classification process constitutes the final stage of our model, where we integrate insights derived from the detection phase with a Multi-Layer Perceptron (MLP) layer to predict the emotional state of the targeted utterance.

We employ a skip connection to concatenate original intra-context embedding $c_{i,s}^{\text{intra}}$ with the output of the cross-verification layer $\hat{c}_{i,s}^{\text{intra}}$ along the feature dimension axis. The concatenated feature vector $\mathbb{F}_{i,s}$ represents the updated embedding of the i-th utterance from speaker s:

$$\mathbb{F}_{i,s} = \text{Concat}\big(\hat{c}_{i,s}^{\text{intra}}, c_{i,s}^{\text{intra}}\big) \quad (5)$$

Next $\mathbb{F}_{i,s}$ is fed into the MLP for further processing. The MLP transforms $\mathbb{F}_{i,s}$ into a high-level representation $h_{i,s}$ for making a final prediction:

$$h_{i,s} = \text{MLP}(\mathbb{F}_{i,s}) \quad (6)$$

In the final step, we employ the softmax function to the output of the MLP layer $h_{i,s}$ to obtain a

probability distribution over the possible emotional states. The predicted emotional state $\hat{y}_{i,s}$ for the targeted utterance is thus given by:

$$\hat{y}_{i,s} = \text{Softmax}(h_{i,s}) \quad (7)$$

4 Experiments and Results

4.1 Datasets

DetectiveNN was tested on three benchmark datasets for recognizing emotions in conversations: IEMOCAP (Busso et al., 2008), EmoryNLP (Zahiri and Choi, 2018), and DailyDialog (Li et al., 2017). While IEMOCAP and DailyDialog are designed for dialogues between two parties, EmoryNLP is designed to learn from multi-party conversations. We report experimental results for conversational emotion recognition for all three datasets. The details of all datasets are presented in Table 1.

IEMOCAP (Busso et al., 2008): IEMOCAP is a dataset of two-person conversations among ten different unique speakers. However, for training purposes, only the first eight speakers from sessions one to four are included. Each video in this collection captures an individual dyadic dialogue, which is further divided into separate utterances. These utterances have been classified with annotations corresponding to six different emotional states: happiness, sadness, neutrality, anger, excitement, and frustration.

EmoryNLP (Zahiri and Choi, 2018): EmoryNLP utilizes content from the TV series "Friends". This dataset includes utterances that are classified into seven specific emotional categories: neutral, joyful, peaceful, powerful, scared, mad, and sad. Additionally, sentiments in this dataset are identified as either positive, negative, or neutral.

DailyDialog (Li et al., 2017): DailyDialog covers a wide array of topics pertinent to everyday life, closely mirroring the style of natural human conversation. This dataset is distinctive in that each of its utterances is annotated with labels for both emotional categories and dialogue acts. It includes a comprehensive range of seven emotional labels: angry, disgusted, fearful, joyful, neutral, sad, and surprised.

Our research primarily investigates the emotional categorization and text aspects of these datasets. We align our study with COSMIC's (Ghosal et al., 2020) train/validation/test splits for consistency.

4.2 Baselines

We compare our model, DetectiveNN, with several models introduced in the related work section, including DialogueRNN, DialogueGCN, DialogueCRN, BC-LSTM, CMN, EmotionIC, COSMIC, and DialogXL. Additionally, we also evaluate DetectiveNN against two other models: EmoCaps and CNN.

EmoCaps (Li et al., 2022): EmoCaps utilizes a transformer-based architecture to extract emotional trends across various modalities. It leverages a bi-directional LSTM for contextual analysis, integrating both past and future conversational context to classify emotions.

CNN (Kim, 2014): CNN is a convolutional neural network designed to be trained on utterances that are context-independent.

Table 2, Table 3, and Table 4 present the performance evaluation of DetectiveNN on the test data. In training the model on the IEMOCAP dataset, we integrate texutal, visual and aduio features to create multimodal fused embeddings. All three modality feature embeddings are obtained from Li et al. (2022). For training the model on the EmoryNLP and DailyDialog datasets, we utilized RoBERTa to extract contextual features. RoBERTa embeddings are taken from Ghosal et al. (2020).

4.3 Evaluation Metrics

Consistent with prior studies by Hazarika et al. (2018a), Ghosal et al. (2020), and Jiao et al. (2020), we select the accuracy score (Acc.) as our primary metric for evaluating overall performance on the IEMOCAP, EmoryNLP, and DailyDialog datasets. Additionally, to provide a comprehensive assessment of our model's capability across both majority and minority classes, we report both the Weighted-average F1 score (Weighted-F1) and the Macro-averaged F1 score (Macro-F1) for IEMOCAP and EmoryNLP datasets. We report both the micro-average F1 score (Micro-F1) and the Macro-averaged F1 score (Macro-F1) for the DailyDialog dataset. These metrics offer a more nuanced view of the model's effectiveness in handling different class distributions.

4.4 Implementation Details

In our experimental setup, the validation set is utilized for hyperparameter optimization. The architecture varies between datasets: a single-layer bidirectional GRU is applied to IEMOCAP, EmoryNLP

Dataset	# Dialogues			# Utterances			Avg.	# Classes
	train	val	test	train	val	test	Length	
IEMOCAP	108	12	31	5,810	—	1,623	47	6
DailyDialog	11,118	1,000	1,000	87,832	7,912	7,863	72	7
EmoryNLP	659	89	79	7,551	954	984	10	7

Table 1: Table 1: The statistics of three datasets.

and Dailydialog datasets. In the subsequent detection phase, a two-layer transformer encoder block is used for the EmoryNLP dataset, while a one-layer transformer encoder block is used for both IEMOCAP and Dailydialog datasets.

The batch size is uniformly maintained at 30 for all experiments. For each dataset, the learning rate and dropout are set specifically as follows: 10^{-3} and 0.5 for IEMOCAP, 10^{-4} and 0.2 for EmoryNLP, and 10^{-4} and 0.5 for Dailydialog. L2 weight decay is set to 2×10^{-3} for all experiments. The loss objective for all experiments is cross-entropy loss. We trained the DetectiveNN for a maximum of 80 epochs using the Adam optimizer (Kingma and Ba, 2014) and stopped training if the validation loss does not decrease for 10 consecutive epochs. For benchmarking against existing models like CNN, BC-LSTM, DialogueGCN, DialogueRNN, and DialogueCRN, we replicate their setups using the publicly available code provided by Kim (2014), Poria et al. (2018), Majumder et al. (2019), Ghosal et al. (2019), and Hu et al. (2021), ensuring consistency in the experimental environment.

Methods	Acc.	Weighted-F1	Macro-F1
CNN †	53.16	52.13	47.28
BC-LSTM †	55.86	55.24	53.19
CMN*	56.56	56.13	54.30
COSMIC*	–	65.28	–
DialogXL*	–	65.94	–
DialogueRNN†	63.50	63.18	62.99
DialogueGCN†	62.42	62.11	61.17
DialogueCRN†	70.65	70.35	70.01
EmoCaps*	–	**71.77**	–
DetectiveNN	**76.15**	**76.01**	**76.40**

Table 2: Experimental results on the IEMOCAP dataset. Annotated with an * indicates results sourced from the model's paper, and a (†) denotes results from reproductions conducted by the authors.

4.5 Main Results

Table 2, Table 3, and Table 4 illustrate the results of comparing our DetectiveNN model with other models and backbones from different perspectives.

Methods	Acc.	Micro-F1	Macro-F1
CNN†	65.35	57.21	50.13
BC-LSTM†	64.19	53.19	48.94
EmotionIC*	–	60.13	54.19
COSMIC*	–	58.48	51.05
DialogXL*	–	54.93	–
DialogueRNN†	63.03	61.50	57.66
DialogueGCN†	71.56	62.20	**60.43**
DialogueCRN†	**73.15**	**64.10**	53.18
DetectiveNN	**75.55**	**70.20**	**57.38**

Table 3: Experimental results on the Dailydialog dataset. Annotated with an * indicates results sourced from the model's paper, and a (†) denotes results from reproductions conducted by the authors.

Methods	Acc.	Weighted-F1	Macro-F1
CNN†	34.21	30.19	28.59
BC-LSTM†	38.17	34.27	29.87
SACL-LSTM*	–	39.65	–
COSMIC*	–	38.11	–
DialogXL*	–	34.73	–
DialogueGCN†	37.75	34.98	31.30
DialogueCRN†	40.65	37.59	32.31
DialogueRNN†	41.04	35.76	31.22
EmotionIC*	–	**40.25**	–
DetectiveNN	**42.68**	**40.78**	**33.65**

Table 4: Experimental results on the EmoryNLP dataset. Annotated with an * indicates results sourced from the model's paper, and a (†) denotes results from reproductions conducted by the authors.

Based on this, we make the following observations:

(1) Our method achieves significant improvements over the SOTA baseline models on all benchmarks. Specifically, we outperform EmoCaps, DialogueCRN, and EmotionIC by 4.24%, 6.10%, and 0.53% on IEMOCAP, Dailydialog and EmoryNLP respectively.

(2) DetectiveNN improves over all the models; however, the performance gain of the model on the IEMOCAP dataset is not as significant as it is on the DailyDialog dataset. DetectiveNN achieves new state-of-the-art scores of 70.20% for Micro-F1 and 75.55% for Accuracy on DailyDialog.

(3) Previous research has highlighted the complexity involved in emotion modeling in the EmoryNLP dataset, challenges stemming from the diversity of speakers and limited conversational

exchanges (Ghosal et al., 2019; Li et al., 2020). DetectiveNN, in contrast, shows notable performance enhancements on the IEMOCAP and DailyDialog datasets. This advancement is attributed to longer and more in-depth conversational exchanges and richer utterance content in these datasets. These aspects allow for a more comprehensive understanding of the global context and emotional cues, thus enhancing the accuracy of DetectiveNN.

4.6 Ablation Study

The DetectiveNN model is built on a recall-detect-predict framework. To understand the impact of its recall and detection phases on overall performance, we conducted a series of ablation experiments on both the IEMOCAP and EmoryNLP datasets. When two modules are removed successively, the performance is greatly decreased. This indicates the importance of both the recall phase and the detection phase in the DetectiveNN model. The outcomes of these experiments are detailed in Table 5.

Recall Phase Analysis: The recall phase plays a crucial role in gathering relevant global context from dialogues. As indicated in the second column of our results, excluding this phase led to a notable reduction in the model's effectiveness on both datasets. This result demonstrates that the essential nature of the recall phase in forming a contextual base, which is crucial for the subsequent reasoning process.

Detection Phase Analysis: In the subsequent set of experiments, we focused on the removal of the detection phase, a critical component for analyzing emotional cues retrieved in the recall phase. The absence of this phase resulted in a marked decrease in performance across both datasets, as highlighted in our results. This decline shows the critical role of the detection phase in decoding emotional cues in a conversational context. Furthermore, our findings, as detailed in the final row of Table 5, reveal that eliminating both the recall and detection phases also results in a significant drop in performance. This marked decline highlights the interdependent and synergistic nature of these two phases, underlining their combined importance in augmenting the reasoning capability of the DetectiveNN model.

Impact of Intra-Contextual Dependency: Our study further explored the significance of intra-contextual dependency, essential for understanding how a speaker's emotional state is shaped by their unique conversational context. Excluding this dependency-tracking component from DetectiveNN resulted in a great decline in performance across both datasets. This outcome highlights the imperative for DetectiveNN to effectively monitor each speaker's emotional journey, allowing the model to accurately identify and interpret personal emotional cues.

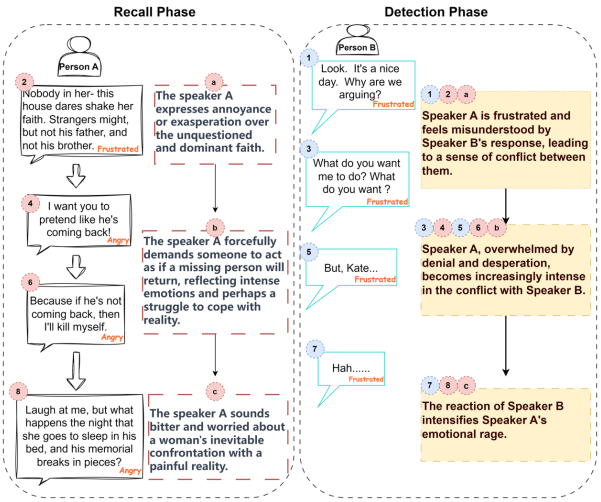

Figure 2: The case study

4.7 Case Study

Conventional methods like DialogueRNN and DialogueGCN encounter difficulties in interpreting emotional cues, such as discerning the root causes and intensity of emotions such as conflict, leading to inaccurate recognition of the emotion as frustrated or neutral. Although DialogueCRN possesses the cognitive ability to retrieve and combine emotional cues from conversational contexts, it cannot further interpret those cues for accurate prediction. In contrast, our model, DetectiveNN, utilizes a recall-detect-predict framework to decipher emotional clues more effectively. Figure 2 shows a conversation sampled from the IEMOCAP dataset. The goal is to recognize the emotional state of the targeted utterance 8 from person A.

DetectiveNN operates through two iterative phases:

- The Recall Phase: This phase is dedicated to extracting relevant emotional cues from a sequence of utterances (2, 4, 6, and 8) made by speaker A, denoted by red circles.
- The Detection Phase: This phase involves examining emotional cues within an emo-

Context	Cognition		IEMOCAP			EmoryNLP		
Intra-Contextual Dependency	**Recall Phase**	**Detection Phase**	**Acc.**	**W-F1**	**M-F1**	**Acc.**	**W-F1**	**M-F1**
✓	✓	✓	**76.15**	**76.01**	**76.40**	**42.68**	**40.78**	**33.65**
✓	×	✓	51.60	50.38	50.62	38.21	36.03	29.15
✓	✓	×	41.46	38.60	36.77	39.11	37.39	31.15
×	✓	✓	70.40	70.68	70.98	40.55	38.35	30.85
×	×	✓	57.74	57.13	57.80	37.60	37.10	30.45
×	×	×	50.26	50.14	50.30	38.92	37.00	30.07

Table 5: Experimental results of ablation studies on IEMOCAP and EmoryNLP datasets.

tional narrative (indicated by red circles a-b-c) against the context of the entire conversation from utterances 1 to 8 (marked by both red and blue circles). It decodes these emotional cues to uncover patterns in the emotional flow of speaker A's responses (indicated by yellow blocks).

This dual-phase process enables the model to make a precise final prediction. DetectiveNN, in this case, identifies an intensifying dispute between speakers A and B. This nuanced understanding of the escalating conflict leads to a more accurate identification of the emotion as anger, rather than an incorrect prediction of it being neutral or depressed.

5 Conclusions

In this paper, we introduce DetectiveNN, a novel framework for Emotion Recognition in Conversation. This framework utilizes an innovative recall-detect-predict structure to interpret emotions in conversations. Initially, DetectiveNN identifies key emotional cues within the dialogue. Subsequently, it conducts a thorough analysis of these cues to accurately predict the emotional state.

Rigorously evaluated across three benchmark datasets, DetectiveNN has demonstrated its superiority over existing models, revealing the profound impact of integrating cognitive reasoning into deep learning architectures. This cognitive factor plays an important role not only in enhancing the model's efficiency and accuracy in prediction but also in advancing ERC methodologies.

6 Limitations

DetectiveNN boosts emotion prediction accuracy through its analysis of long-term dialogue turns but faces challenges with short-term turns due to its dependence on extended interaction context. This dependence constrains its capability to identify and understand emotional cues in brief dialogic exchanges. Moreover, the lack of information on a speaker's personality traits impacts DetectiveNN's ability to capture complex emotional dynamics, as seen in datasets like EmoryNLP from the "Friends" TV series. Speakers' personality traits are essential for identifying sarcasm, humor, and other subtle emotional cues. Without integrating this knowledge, DetectiveNN struggles to accurately predict nuanced emotions in conversations.

References

Lisa Feldman Barrett. 2014. The conceptual act theory: A précis. *Emotion review*, 6(4):292–297.

Lisa Feldman Barrett and James A Russell. 2014. *The psychological construction of emotion*. Guilford Publications.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Shauna Concannon and Marcus Tomalin. 2023. Measuring perceived empathy in dialogue systems. *AI & SOCIETY*, pages 1–15.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network

for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

M Gomathy. 2021. Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. *International Journal of Speech Technology*, 24(1):155–163.

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

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

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 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)*, pages 7042–7052, Online. Association for Computational Linguistics.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *INTERSPEECH*, pages 2347–2351.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

James A Russell. 2009. Emotion, core affect, and psychological construction. *Cognition and emotion*, 23(7):1259–1283.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Haiqin Yang and Jianping Shen. 2021. Emotion dynamics modeling via bert. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Liu Yingjian, Li Jiang, Wang Xiaoping, and Zeng Zhigang. 2023. Emotionic: Emotional inertia and contagion-driven dependency modelling for emotion recognition in conversation. *arXiv preprint arXiv:2303.11117*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.