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

Simin Hong
Zhejiang Lab
Hangzhou, China
cliosimin@zhejianglab.org

Jun Sun *
Zhejiang Lab
Hangzhou, China
sunjun16sj@gmail.com

Taihao Li
Hangzhou Institute for Advanced Study, UCAS
Hangzhou, China
lith@ucas.ac.cn

Abstract

Emotion Recognition in conversations (ERC) 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 since they are insufficient in understanding the rich historical emotional context. 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 current emotional state. Tested on three benchmark datasets, our approach significantly outperforms existing methods. This highlights the advantages of incorporating cognitive factors into deep learning for ERC, enhancing task efficacy 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 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; Yang et al., 2024).

In the realm of conversational emotion recognition, interpreting emotional cues embedded in conversational context is crucial (Mittal et al., 2020; Gomathy, 2021). Emotional cues in conversations are subtle patterns or indicators that hint at the underlying emotions of a speaker. These cues often act as triggers for the emotions expressed in current utterances (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 an 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 combines emotional clues from the dialogue. However, these methods lack a distinct process for interpreting extracted emotional cues before classifying the emotion.

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 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 involves identifying and organizing emotional cues,

*Corresponding author:Jun Sun (sunjun16sj@gmail.com).
[1]our code can be found here

Findings of the Association for Computational Linguistics: EACL 2024, pages 9170–9180
November 12-16, 2024

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 Gated Recurrent Units (GRUs), a type of sequence-based model, for their efficiency in handling sequential data and their ability to capture long-term dependencies. This is crucial for retaining diverse contextual information from the speakers' emotional experiences and interactions. 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. During the examination phase, we employ transformer encoder blocks with Multi-Head Attention (MHA) layers. MHA layers enable the model to simultaneously focus on various parts of the input sequence by assigning disparate levels of attention or importance. Transformer encoder blocks integrate cues to achieve a deep understanding of the speaker's emotional context. 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 advanced significantly, emphasizing the extraction and integration of emotional cues 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) uses GRUs to track emotional states by integrating speaker identity, context, and emotions from neighboring utterances. DialogueCRN (Hu et al., 2021) combines cognitive theories of emotion with LSTM networks for iterative extraction and integration of emotional cues. BC-LSTM (Poria et al., 2018) employs bidirectional LSTMs to capture the influence of preceding and following utterances. CMN (Hazarika et al., 2018b) models past utterances using GRUs in a multimodal approach. EmotionIC (Yingjian et al., 2023) inte-

grates attention and recurrence with IMMHA, DiaGRU, and SkipCRF for comprehensive emotion detection. COSMIC (Ghosal et al., 2020) incorporates commonsense knowledge with GRUs to model various conversational states. SACL-LSTM (Hu et al., 2023) uses contrastive learning to compare emotional representations from both original and adversarial data, allowing the model to better generalize in recognizing emotions across different conversational contexts, while DF-ERC (Li et al., 2023) uses contrastive learning to separate modality and utterance features. EACL (Yu et al., 2024) utilizes label encodings as anchors and develops an auxiliary loss to better distinguish similar emotions.

Pre-trained Language Model-based Models: DialogXL (Shen et al., 2021) uses XLNet (Yang et al., 2019) and dialogue-level self-attention to handle multi-party conversation dynamics. Emoberta (Kim and Vossen, 2021) employs RoBERTa (Liu et al., 2019) to predict speaker emotions by learning both speaker-level and dialogue-level context.

Graph-based Models: DialogueGCN (Ghosal et al., 2019) uses a graph convolutional neural network to model conversational context by representing utterances as nodes and their dependencies as edges. Zhang et al. (2019) and Lian et al. (2020) employ graph convolutional networks with attention mechanisms to capture context-sensitive and speaker-sensitive dependencies.

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.

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.

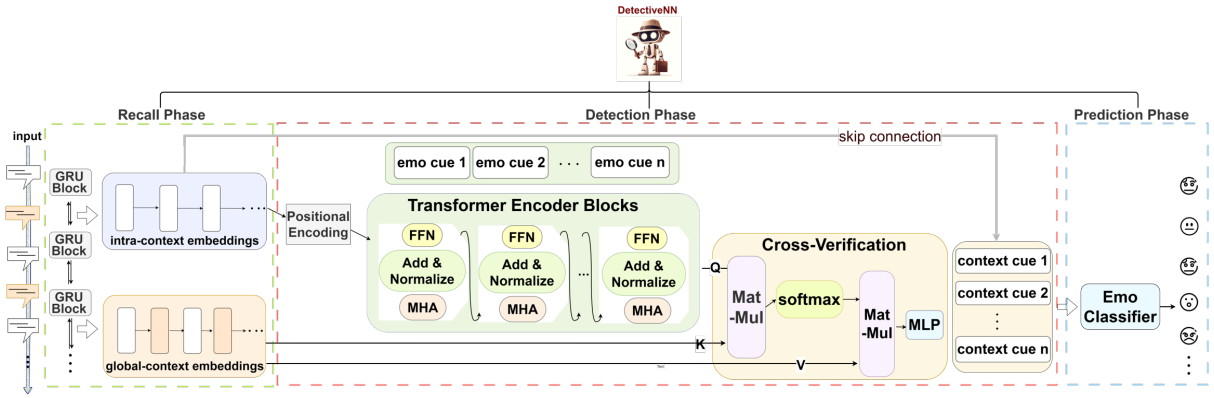

Figure 1: The architecture of the proposed model DetectiveNN

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 emotional cues from intra-context 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 $\tilde{C}_s^{\text{intra}}$ 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). IEMOCAP and DailyDialog focus on two-party dialogues, while EmoryNLP includes multi-party conversations. We report results for all three datasets, with details in Table 1.

IEMOCAP (Busso et al., 2008): IEMOCAP consists of two-person conversations among ten speakers, with training data from the first eight speakers. Each video captures a dyadic dialogue, divided into utterances annotated with six emotions: happiness, sadness, neutrality, anger, excitement, and frustration.

EmoryNLP (Zahiri and Choi, 2018): EmoryNLP utilizes content from the TV series "Friends," this dataset includes utterances classified into seven emotions: neutral, joyful, peaceful, powerful, scared, mad, and sad. Sentiments are labeled as positive, negative, or neutral.

DailyDialog (Li et al., 2017): DailyDialog covers various everyday topics, mirroring natural human conversation. Each utterance is annotated with emotional categories and dialogue acts, including seven emotions: 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, DialogXL, EACL, SACL-LSTM and DF-ERC. 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 utilize 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 and Dailydialog datasets.

In the subsequent detection phase, we employ two transformer encoder blocks because the EmoryNLP dataset is characterized by a limited number of turns and brief conversations. This configuration facilitates a more prolonged learning process, allowing the model to effectively detect nuanced emotional cues within short conversational contexts. Our experiments also demonstrate that incorporating additional encoder blocks enables the model to identify a broader range of features. For the IEMOCAP and Dailydialog datasets, which are

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.

characterized by longer turns and more extended conversations, we implement a single transformer encoder block.

The batch size is uniformly set to 30 for all experiments. The initial hyperparameter search space includes learning rates of 10^{-3}, 10^{-4}, and 10^{-5}, and dropout rates of 0.2, 0.3, 0.4, and 0.5. Based on the results of the grid search, the selected configurations for each dataset are 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 train the DetectiveNN for a maximum of 80 epochs using the Adam optimizer (Kingma and Ba, 2014) and stop 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
Emoberta*	–	68.57	–
EACL*	68.81	70.41	–
SACL-LSTM*	69.08	69.22	–
DF-ERC*	71.84	71.75	–
EmotionIC*	69.44	69.61	–
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.

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
SACL-LSTM*	42.21	39.65	–
EACL*	36.45	40.24	–
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.

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. 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) 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 based on a recall-detect-predict framework. To assess the impact of the recall and detection phases on performance, ablation experiments were conducted on the IEMOCAP and EmoryNLP datasets. Removing either phase successively led to a significant performance drop, demonstrating their importance. Detailed results are in Table 5.

Recall Phase Analysis: The recall phase gathers relevant global context from dialogues. Excluding this phase reduced the model's effectiveness on both datasets, highlighting its crucial role in forming a contextual base for reasoning.

Detection Phase Analysis: The detection phase analyzes emotional cues derived from the recall phase. Notably, when the recall phase was removed, transformer encoders were applied to both intra-context and global context inputs. As highlighted in our results, the absence of the detection phase resulted in a marked decrease in performance across both datasets, indicating its critical role in decoding emotional cues within a conversational context. Furthermore, our findings, as detailed in the final row of Table 5, reveal that eliminating both the recall and detection phases together results in a significant drop in performance. This marked decline underscores the interdependent and synergistic nature of these two phases, underlining their combined importance in enhancing the reasoning capability of the DetectiveNN model.

Impact of Intra-Contextual Dependency: Our study further explores 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 results 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.

4.7 Comparative Case Study

We also conduct a comparative case study to evaluate our method against the DialogueCRN and DialogueRNN models. Table 6 presents a conversation sampled from the IEMOCAP dataset. DialogueRNN fails to capture the complex emotional context in certain utterances, such as mislabeling "I want you to pretend like he's coming back!" as "sad," thereby missing the underlying frustration and possible anger. Similarly, it incorrectly predicts "excited" for "But, Kate...," indicating a lack of understanding of emotional nuances and situational context. DialogueCRN also demonstrates limitations, such as misinterpreting conflict by incorrectly labeling the utterance "Laugh at me, but what happens the night that she goes to sleep in his bed, and his memorial breaks into pieces?" as "sad."

In contrast, our model, DetectiveNN, employs a recall-detect-predict framework that demonstrates more accurate emotion recognition in dialogues. In this case, the predicted labels from DetectiveNN alternate between "frustrated" and "angry," while the predictions from DialogueCRN and DialogueRNN exhibit more varied and less stable labels. The results suggest that DetectiveNN takes better advantage of historical information, meaning that it respects emotional inertia.

5 Visualization & Analysis

We present heatmap visualizations to interpret the model's emotion predictions using a randomly selected conversation sample from the IEMOCAP test dataset. The first heatmap (see Figure 2) illustrates features extracted from the transformer encoder's final layer during the detection phase, while the second heatmap (see Figure 3) shows activations from the first layer of the MLP. In both visualizations, the x-axis represents the dimensions of the feature vector, each value corresponding to the average feature activation for a specific emotion class. The y-axis denotes different emotion classes.

The first heatmap reveals distinct activation patterns across the 200 feature dimensions, which vary significantly between emotion classes such as ***angry*** and ***sad***. These activation patterns may capture lexical features indicative of specific emotional cues. However, the model struggles to distinguish

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.

Ground-truth label	Detective NN	Dialogue CRN	Dialogue RNN	A Case Study
frustrated	frustrated	frustrated	neutral	Person B: Look. It's a nice day. Why are we arguing?
frustrated	frustrated	angry	angry	Person A: Nobody in her—this house dares shake her faith. Strangers might, but not his father, and not his brother.
frustrated	frustrated	neutral	frustrated	Person B: What do you want me to do? What do you want–
angry	angry	frustrated	sad	Person A: I want you to pretend like he's coming back!
frustrated	frustrated	neutral	excited	Person B: But, Kate...
angry	angry	excited	angry	Person A: Because if he's not coming back, then I'll kill myself.
frustrated	frustrated	frustrated	excited	Person B: Hah.....
angry	angry	sad	angry	Person A: Laugh at me, but what happens the night that she goes to sleep in his bed, and his memorial breaks in pieces?

Table 6: A comparative case study.

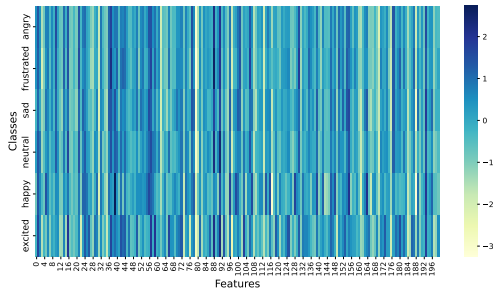

Figure 2: Output of the last layer of the transformer encoder block.

Emotion	angry	frustrated	sad	neutral	happy	excited
angry	1.00	**0.98**	0.30	0.96	0.26	0.41
frustrated	0.98	1.00	0.43	0.99	0.39	0.54
sad	0.30	0.43	1.00	0.46	0.97	0.90
neutral	0.96	0.99	0.46	1.00	0.45	0.61
happy	0.26	0.39	0.97	0.45	1.00	0.96
excited	0.41	0.54	0.90	0.61	0.96	1.00

Table 7: Emotion similarity matrix for emotion classes based on the output of the last layer of the transformer encoder block.

closely related negative emotions, such as ***angry*** and ***frustrated***. This difficulty suggests that emotional cues alone may not be sufficient to capture fine-grained differences between similar emotions.

To further validate these observations, We compute a cosine similarity matrix using the average activations from the output of the final layer of the transformer encoder across the entire test dataset for each emotion class (see Table 7). The results numerically confirm our findings from Figure 2, reinforcing the model's difficulty in distinguishing similar emotions, particularly among negative emotions, as indicated by high similarity scores.

The second heatmap (see Figure 3) shows more intense activations for the ***frustrated*** class compared to the ***angry*** class, suggesting that the model integrates contextual information to decode the complex interplay between emotions, events, and the unfolding narrative within a sentence. To quantify these activations, we calculate a cosine similarity matrix based on the average activations from the output of the first layer of the MLP for each emotion class across the entire test dataset (see Table 9). We then compare these results with the average activations from the DetectiveNN model without the detection phase (see Table 8). To ensure a fair comparison, we increase the number of MLP layers to match the parameter count of the original model.

Table 9 demonstrates the significant impact of the detection phase on the model's performance. When this phase is removed, the similarity between closely related emotions, such as ***angry*** and ***frus-***

Figure 3: Output of the first layer of MLP.

Emotion	angry	frustrated	sad	neutral	happy	excited
angry	1.00	**0.83**	-0.02	0.28	-0.82	-0.74
frustrated	0.83	1.00	0.19	0.76	-0.59	-0.39
sad	-0.02	0.19	1.00	0.29	0.35	-0.06
neutral	0.28	0.76	0.29	1.00	-0.06	0.25
happy	-0.82	-0.59	0.35	-0.06	1.00	**0.79**
excited	-0.74	-0.39	-0.06	0.25	0.79	1.00

Table 8: Emotion similarity matrix for emotion classes based on the output of the first layer of MLP without the detection phase.

Emotion	angry	frustrated	sad	neutral	happy	excited
angry	1.00	**0.69**	-0.11	0.18	-0.75	-0.38
frustrated	0.69	1.00	0.21	0.53	-0.74	-0.36
sad	-0.11	0.21	1.00	0.10	-0.18	-0.65
neutral	0.18	0.53	0.10	1.00	-0.34	-0.03
happy	-0.75	-0.74	-0.18	-0.34	1.00	**0.64**
excited	-0.38	-0.36	-0.65	-0.03	0.64	1.00

Table 9: Emotion similarity matrix for emotion classes based on the output of the first layer of MLP with the detection phase.

trated (0.83 vs. 0.69), as well as ***happy*** and ***excited*** (0.79 vs. 0.64). These results indicate a higher difficulty in distinguishing similar emotions when the detection phase is absent. Conversely, the model's ability to differentiate broader emotion categories improves, reinforcing the conclusion that contextual information is effectively captured and integrated across different model layers.

6 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.

7 Limitations

DetectiveNN improves emotion prediction accuracy for long-term dialogue turns but struggles with short-term turns due to its reliance on extended interaction context. This context dependence limits its ability to detect emotional cues in brief exchanges. Additionally, lacking information on speakers' personality traits hinders DetectiveNN's performance in capturing complex emotional dynamics, such as sarcasm and humor, which are prevalent in datasets like EmoryNLP from the "Friends" TV series. Integrating personality trait knowledge is essential for accurately predicting nuanced emotions in conversations.

Acknowledgments

The work by Jun Sun is supported by the National Natural Science Foundation of China (Grant No. 62306289).

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, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

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*.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934.

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 multiparty 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.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*.

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*.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. *arXiv preprint arXiv:2403.20289*.

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*.