Received 12 December 2023, accepted 23 December 2023, date of publication 1 January 2024,
date of current version 8 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348518

RESEARCH ARTICLE

Residual Relation-Aware Attention Deep Graph-Recurrent Model for Emotion Recognition in Conversation

ANH-QUANG DUONG, NGOC-HUYNH HO, (Member, IEEE), SUDARSHAN PANT, (Member, IEEE), SEUNGWON KIM, SOO-HYUNG KIM, (Member, IEEE), AND HYUNG-JEONG YANG, (Member, IEEE)
Department of Artificial Intelligent Convergence, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government through the Ministry of Science and Information and Communications Technology (MSIT), South Korea, under Grant RS-2023-00219107; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development Grant funded by the Korea Government (MSIT) under Grant IITP-2023-RS-2023-00256629; and in part by the MSIT under the Innovative Human Resource Development for Local Intellectualization Support Program supervised by the IITP under Grant IITP-2023-RS-2022-00156287.

ABSTRACT This work addresses Emotion Recognition in Conversation (ERC), a task with substantial implications for the classification of the underlying emotions in spoken encounters. Our focus is on utilizing a fully connected directed acyclic graph to represent conversations, presenting inter-locutor and intra-locutor ties to capture intricate relationships. Therefore, we propose a novel methodology, Residual Relation-Aware Attention (RRAA) with Positional Encoding, enhancing speaker relations' contexts for improved emotion recognition in conversation. The purpose of this mechanism is to facilitate a thorough comprehension of the connections between speakers, hence enhancing the sophistication and contextual awareness of an emotion recognition framework. We utilized the Gated recurrent units (GRU) to regulate context transmission, ensuring adaptability to changing emotional dynamics. It regulates the transmission of conversation context across all layers of the graph, guaranteeing a flexible and responsive representation of the changing emotional dynamics within the discourse. Evaluations on IEMOCAP, MELD, and EmoryNLP datasets disclose our model's superior performance (F1 scores: 69.1%, 63.82%, 39.85%, respectively), outperforming state-of-the-art approaches. In general, this work enhances speaker interactions by utilizing a fully connected graph, and providing a more concise and efficient ERC framework.

INDEX TERMS Emotion recognition in conversation, residual relation-aware attention, deep graph-recurrent model, fully connected directed acyclic graph, positional encoding.

I. INTRODUCTION

The topic of emotion recognition in conversation has gained considerable attention due to the increasing prevalence of open discussions on many platforms, such as social media and film franchises. The motivation behind this fascination is driven by the inherent complexity of interpreting the emotional aspects of a spoken statement, which is influenced by various elements like the topic being discussed, the speaker's viewpoint, and the unique characteristics of their personality. Moreover, the emotional tone attributed to a particular segment of speech is subject to dynamic modification through following utterances, whether they originate from the same speaker or different speakers. To tackle this challenge, it is crucial to effectively model the context of the conversation. A strong representation of the context greatly improves the performance of the model, especially when individual utterances do not provide sufficient information

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

for independent recognition, which is demonstrated by the DialogueGCN model [1].

Nowadays, there are three prevailing strategies utilized for modeling conversation context: graph-based methods [1], [2], recurrent-based methods [3], and transformer-based methods [4]. Graph-based methodologies [1], [2] demonstrate exceptional proficiency in capturing relation-aware characteristics among several participants in a conversational context. However, numerous research endeavors commonly utilize predetermined window sizes, which might have a detrimental effect on the feasibility of using these techniques in real-time applications. In contrast, recurrent-based techniques [3] prioritize sequential processing, allowing them to effectively capture temporal dependencies in conversations. Although these strategies can be helpful in specific situations, they may fail to consider complex multi-party relationships. Transformer-based approaches, such as TODKAT [4], utilize attention mechanisms to improve computing efficiency. Although transformer-based models have benefits in parallel processing, they may accidentally ignore sequential information that is essential to understanding conversation dynamics. In this context, we refer to the DAGERC framework [5] that utilizes a directed acyclic graph to represent conversational interactions, with a specific emphasis on capturing the influence of prior utterances on future ones. Nevertheless, the model's capacity to fully use linkages among speakers is constrained by restricted connections.

In response to these considerations, our research endeavors to present an innovative methodology derived from DAGERC, which aims to predict emotional states in utterances inside conversational contexts. The main focus of our research is the efficient exploitation of a fully connected directed acyclic graph for the purpose of modeling talks. This process entails the establishment of links between each individual node and its preceding nodes, so providing a nuanced depiction of the dynamics within a conversation. In order to incorporate relation-aware features, we provide a novel approach called Residual Relation-aware Attention (RRAA), which is enhanced by the inclusion of Positional Encoding (PE). This allows for the consideration of sequence information, particularly in graphs that exhibit a high number of connections. Furthermore, the integration of gated recurrent units (GRUs) [6] is employed to regulate the transmission of conversational context between graph-recurrent layers. Our experiments, conducted on three diverse ERC datasets, demonstrate the efficacy of our proposed model, surpassing the performance of state-of-the-art methods. The central goal of our research is to enrich inter-speaker and intra-speaker relations through a fully connected directed acyclic graph, overcoming limitations in the number of connections among speakers evident in existing studies like DialogueGCN and DAGERC.

The main contributions of our work can be summarized as follows: (1) We employ a fully connected directed acyclic graph with two types of edge relations to comprehensively leverage relations among speakers. (2) We introduce residual relation-aware attention with positional encoding and GRUs to capture relation-aware features and control conversation context propagation through graph-recurrent layers. (3) Experimental results showcase improvements over state-of-the-art methods on benchmark ERC datasets, validating the effectiveness of our proposed methodology.

The subsequent sections of the paper are organized as follows: Section II reviews related works on recurrent-based methods, transformer-based methods, and graph-based methods in the ERC task. Section III provides a problem statement and our objectives. Section IV a detailed description of our proposed method. Section V presents comprehensive data information, implementation setup, and experiment results compared with other methods. Finally, Section VI concludes the paper and outlines avenues for future research.

II. RELATED WORKS

In this extended exploration of methods for emotion recognition in conversation, we delve into recent references that contribute to the evolving landscape of this field, encompassing recurrent-based methods, transformer-based methods, graph-based methods, and the specialized domain of directed acyclic graph neural networks.

A. EMOTION RECOGNITION IN CONVERSATION

The comprehension of an individual's emotional condition, which is based on a synthesis of cognitive processes, affective experiences, and behavioral manifestations, may be traced back to Charles Darwin's evolutionary account of emotions during the latter part of the 19th century [7]. Plutchik's seminal work on emotions, as outlined in his 1984 publication [8], established a comprehensive framework for categorizing emotions into eight major kinds. This classification system has since served as a foundational basis for investigating the intricate subtleties and complexities inherent in human emotional experiences. The acknowledgement of the significant influence of language as a reflection of an individual's mental state has led to the widespread adoption of emotion recognition in the field of Natural Language Processing [9], [10]. Strapparava and Mihalcea [11] investigate the emotion recognition in news headlines using the "Affective Text" task. Their goal is to comprehend the relationship between lexical semantics and emotions. Their contribution involves providing a detailed description of the dataset used for evaluation and demonstrating the outcomes of different automated methods for emotion recognition. Strapparava [12] presents the WORDNET-AFFECT resource, which is a lexical representation of affective knowledge derived from WORDNET. This resource provides a supplementary classification of "affective domain labels". It assigns affective meanings to synsets, which are groups of words representing emotional concepts. In addition, Mohammad and Turney [13] addresses the limitation of small emotion lexicons by creating a high-quality, moderate-sized emotion lexicon using Mechanical Turk. Their approach involves implementing word choice

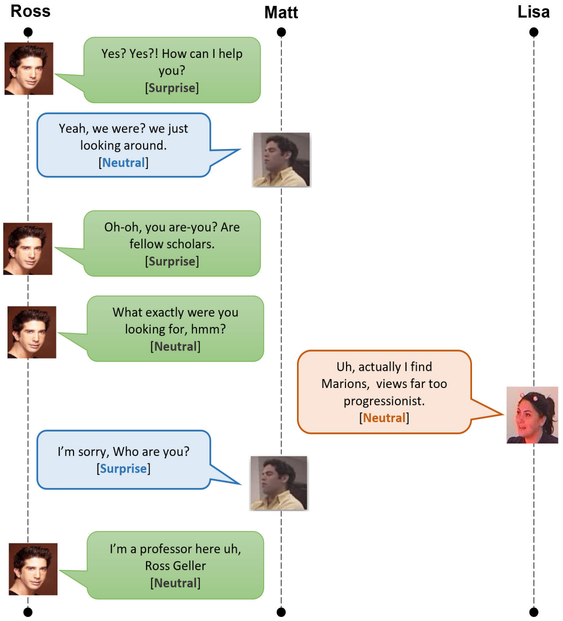

Ross
Matt
Lisa
Yes? Yes?! How can I help you? [Surprise]
Yeah, we were? we just looking around. [Neutral]
Oh-oh, you are-you? Are fellow scholars. [Surprise]
What exactly were you looking for, hmm? [Neutral]
Uh, actually I find Marions, views far too progressionist. [Neutral]
I'm sorry, Who are you? [Surprise]
I'm a professor here uh, Ross Geller [Neutral]

FIGURE 1. Illustration of three-party conversation (Ross, Matt, Lisa). Where different color boxes represent different speakers. (This example is a generated sample not a real sample from a dataset).

problems to discourage malicious entries, detecting instances of unfamiliarity, and acquiring sense-level annotations.

The field of Emotion Recognition in Conversation (ERC) has gained significant attention in scientific research because to the increasing availability of open conversation data from platforms such as social media and movie series [14], [15], [16], [17]. The ramifications of ERC encompass various domains such as chatbots, social networking, and customer service, so contributing to the advancement of our comprehension of interpersonal communication. The importance of incorporating context modeling into individual utterances has been highlighted in research [13]. These studies underline the significance of considering preceding utterances and temporal sequence when doing this modeling procedure. Figure 1 depicts a snapshot of a conversation between three speakers.

B. CONVENTIONAL METHODS FOR ERC

In the realm of recurrent-based methodologies, the work by Jiao et al. [18] presents HiGRU, a hierarchical framework based on Gated Recurrent Units (GRUs). This framework consists of two separate GRU components: a lower-level GRU responsible for modeling inputs at the word level, and an upper-level GRU designed to capture the contextual intricacies present in embeddings at the utterance level. The paradigm introduced by Majumder et al. [19] is extended in DialogueRNN, which utilizes multiple GRUs to capture various conversational contexts such as global state and speaker-listener state. This contextual information is then utilized for the purpose of emotion classification. In a similar manner, the COSMIC model [3] integrates external commonsense knowledge to enhance its performance within the recurrent-based framework.

In the domain of transformer-based methodologies, Li et al. [20] employ a three-block transformer architecture to comprehensively model various facets of conversation, encompassing conversation context, inter-speaker dynamics, and intra-speaker intricacies. EmoBERTa [21] enhances the performance of RoBERTa [22] in the Emotion Recognition in Conversation (ERC) task by introducing speaker names to utterances and embedding separation tokens between utterances in dialogue. TODKAT [4], another representative transformer-based model, integrates external commonsense knowledge of emotion detection in dialogues into its Transformer Encoder-Decoder structure.

The use of graph-based techniques, such as DialogueGCN [1], involves the representation of dialogues as graphs. In this representation, each utterance is connected to the preceding and following utterances. Nevertheless, the lack of feasibility of this method in real-time scenarios, when future speech cannot impact previous utterances and vice versa, necessitates the development of alternative solutions such as RGAT [2]. The proposed research study, conducted by RGAT, integrates positional encoding into the existing DialogueGCN framework. This integration involves the incorporation of relation graph attention mechanism, which facilitates the aggregation of information from adjacent nodes. The ConGCN model [23] employs a representation that considers utterances and speakers as nodes. The edges in this representation capture dependencies that are sensitive to either the speaker or the context. The study conducted by Shen et al. [24] introduces DialogXL, an improved version of XLNET [25], which incorporates sophisticated memory processes and dialog-aware self-attention. The Speaker and Position-aware Graph Neural Network (GNN) model, introduced by Liang et al. [26] under the name S+PAGE, serves as a pioneering approach in the field. This model is specifically tailored to incorporate inter-speaker and intra-speaker contextual dynamics into conversational graphs.

The Directed Acyclic Graph (DAG) is a fundamental structure in neural networks such as DAGRNN [27], DAGNN [28], and DAGERC [5]. DAGRNN addresses the challenge of combining extensive contextual information into localized representations for image elements. This technology enables the computation of DAG-structured images, allowing the network to accurately represent and analyze the relationships between image units that are far apart in a meaningful way. In contrast to its predecessors, DAGNN introduces the capability to stack multiple levels, so enabling each node to gather information from its counterparts on the same layer rather than the preceding layer. The unique architectural design utilizes graph attention mechanisms to aggregate information. DAGERC is an extension of DAGNN that aims to enhance and customize the architecture for the specific job of emotion recognition during conversations, by incorporating the evolutionary progression within directed acyclic graphs.

III. PROBLEM STATEMENT AND OBJECTIVES

The complexity of emotion recognition in conversation arises from the intricate interplay of contextual factors, including the topic under discussion, speaker viewpoints, and individual personality traits. Existing methodologies, encompassing graph-based, recurrent-based, and transformer-based approaches, demonstrate varying strengths and limitations. Graph-based methods excel in capturing relation-aware features but often face challenges in real-time applications due to fixed window sizes. Recurrent-based methods prioritize sequential processing but may neglect multi-party relationships. Transformer-based methods leverage attention mechanisms for efficiency but may overlook sequential information. Moreover, current directed acyclic graph models, while addressing some challenges, have limitations in fully leveraging inter-speaker relations.

Therefore, the primary objective of this paper is to propose an advanced methodology for Emotion Recognition in Conversation (ERC) by building upon the directed acyclic graph architecture. The key goals include:

1) **Efficient Context Modeling**: Develop a methodology that efficiently models conversation context, considering the dynamics influenced by preceding and subsequent utterances within a fully connected directed acyclic graph.
2) **Enhanced Relation-Aware Features**: Introduce Residual Relation-aware Attention with Positional Encoding to capture nuanced relation-aware features, addressing the limitations of fixed window sizes and ensuring effective use of sequence information.
3) **Sequential Information Integration**: Utilize Gated Recurrent Units (GRUs) to regulate the propagation of conversation context through graph-recurrent layers, ensuring the effective integration of sequential information.
4) **Enriched Speaker Relations**: Enrich inter-speaker and intra-speaker relations within the conversation by leveraging the fully connected directed acyclic graph, overcoming limitations observed in existing studies such as DialogueGCN and DAGERC.
5) **Comparative Evaluation**: Conduct extensive experiments on three diverse ERC datasets to evaluate the proposed methodology's performance against state-of-the-art models, demonstrating its efficacy in enhancing emotion recognition accuracy.

By addressing these objectives, this paper aims to contribute to the advancement of ERC methodologies, providing a more nuanced and effective approach for emotion recognition within conversational contexts.

IV. METHODOLOGY

A. FULLY CONNECTED DAG BUILDING FROM CONVERSATION

In the domain of ERC, the structured nature of dialogues is characterized by sequences of utterances denoted as $\{u_1, u_2, \ldots, u_n\}$, each corresponding to emotion labels $\{y_1, y_2, \ldots, y_n\}$. The primary objective of ERC involves predicting accurate emotion labels for individual utterances based on the encompassing conversation context $\{u_1, u_2, \ldots, u_n\}$ and relevant speaker information.

In order to efficiently tackle this challenge, we have developed a conceived and created a Directed Acyclic Graph (DAG) as a modeling framework for the conversational dynamics. The graph, denoted as $G = \{V, E\}$, consists of nodes V that represent individual utterances inside the discussion. The edges E represent the directional connections and relationships between two nodes. Specifically, E can be defined as a tuple (D, R), where D denotes the nodes connected by the edge and R represents the set of edge types, denoted as $R = \{R_1, R_2\}$. The development of connections is governed by a crucial temporal constraint, which ensures that links are generated in accordance with the temporal order of utterances. It is worth noting that every statement is linked to previous utterances, enabling the transmission of information from earlier to later utterances, but preventing the opposite direction.

In the present graph structure, we provide two separate categories of relationships: intra-speaker relationships, where interconnected nodes reflect utterances originating from the same speaker, and inter-speaker relationships, where linked nodes correspond to utterances from different speakers. The distinction stems from the acknowledgment that the emotional effect of a statement is shaped not just by prior statements but also by the speaker who delivers the statement. Hence, the emotional impact originating from a single speaker and that originating from multiple speakers are seen as distinct components that contribute to the emotional context of a spoken statement within a conversation.

Figure 2 illustrates a fully connected DAG model for a three-party conversation, where each node establishes connections with its preceding nodes. Unlike the constraints imposed by DAGERC [5], our model does not restrict the connections of each node to a fixed number of previous nodes. This lack of restriction is particularly beneficial given the inherent variability in conversational data. For instance, as illustrated in the sample Directed Acyclic Graph (DAG) in Figure 2, the connectivity constraints imposed by DAGERC limit the connections of each node to the closest node spoken by the same speaker. Consequently, node 4 is exclusively connected to node 3, with no linkage to nodes 1 or 2. This restriction impedes the model's ability to learn relationships between these nodes, underscoring the importance of adopting a fully connected DAG model to comprehensively leverage relations among speakers. Additionally, we augment the graph with positional encoding for each node, enabling the model to discern the significance of individual nodes within the sequence.

B. DEEP GRAPH-RECURRENT MODEL

The comprehensive structure of the Deep Graph-Recurrent model designed for Emotion Recognition in Conversation (ERC) is visually depicted in Figure 3. In the initial

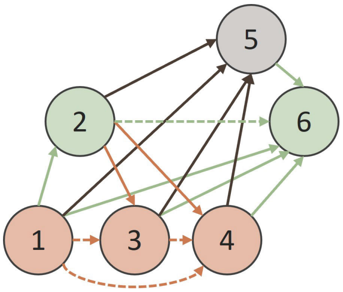


FIGURE 2. DAG of three-party conversation. Each speaker is colored in one of the colors: pale orange, green, or gray. The arrow represents the information propagation from prior nodes to later nodes. Solid lines represent the inter-speaker relation, and the dash lines represent the intra-speaker relation. The number from 1 to 6 represents the order of utterances in dialog.

phase, individual utterances undergo embedding utilizing the RoBERTa model. Subsequently, a graph is systematically generated for each conversational interaction, employing the innovative Fully Connected (DAG) approach. Following the graph construction, a specialized module, comprising residual relation-aware attention with positional encoding and GRUs, is employed. This strategic utilization is intended to adeptly capture relation-aware features while concurrently regulating the propagation of conversational context throughout the layers of the graph-recurrent structure. Ultimately, the outputs derived from all graph-recurrent layers are concatenated, constituting the conclusive representation that informs the final decision-making process.

1) UTTERANCE FEATURE EXTRACTION

Our research utilizes a pre-trained Transformer-based model, RoBERTa, in a way similar to the approaches applied in COSMIC and DAGERC. RoBERTa, also known as Robustly optimized BERT method, represents an enhanced iteration of the Bidirectional Encoder Representations from Transformers (BERT) model. RoBERTa, an AI model developed by Facebook, enhances the original BERT architecture by making improvements to the training objectives and dynamically modifying hyperparameters. These enhancements result in enhanced performance across several natural language processing tasks.

The incorporation of RoBERTa into our system entails the retrieval of feature vectors from utterances. The aforementioned procedure involves the refinement of the model on every Emotion Recognition in Conversation (ERC) dataset. The generated utterance feature representation is produced from the pooled embeddings located in the final layer of the RoBERTa architecture in the fine-tuned model. The feature representation in question possesses a dimensionality of 1024, effectively capturing the intricate linguistic nuances inherent in the utterances. Consequently, it enables a thorough comprehension of the emotional subtleties that exist within conversational situations.

2) RESIDUAL RELATION-AWARE ATTENTION

We introduce a novel mechanism termed residual relation-aware attention (RRAA), designed to systematically process each node within the directed acyclic graph (DAG) by adhering to the established partial order. Specifically, for a given node denoted as u situated at the L^{th} layer of the graph, the resultant output M_u^L is computed through the application of RRAA layer. This computation involves the aggregation of pertinent information from its predecessors denoted as v at the same L^{th} layer, subsequently combining this aggregated information with the intrinsic information pertaining to node u positioned at the $(L-1)^{th}$ layer. To this end, M_u^L is expressed as follows:

$$\begin{aligned} M_u^L &= RRAA^L\left(H_u^{L-1}, H_v^L\right) \\ &= \sum_{v \in P(u)} a_{uv}^L (W_{r1}^L H_v^L Mask1 + W_{r2}^L H_v^L Mask2) + H_u^{L-1} \end{aligned} \tag{1}$$

where the parameters W_{r1}^L and W_{r2}^L, constituting trainable elements, are indicative of two distinct categories of relations inherent in the edge structure, specifically denoting intra-speaker and inter-speaker relations. The utilization of $Mask1$ and $Mask2$ serves the purpose of discerning the nature of the relation involved. The weighting coefficient a_{uv}^L adheres to the design principles of the query-key paradigm within the attention mechanism. Herein, the query element corresponds to the representation of the primary node u within the $(L-1)^{th}$ layer, while v signifies the antecedents of node u. The introduction of a residual connection is imperative to preserve the informational content of the principal utterance, which undergoes partial loss subsequent to the computation of the attention score with its antecedent. Additionally, considering that the emotional characteristics of an utterance are predominantly derived from its own intrinsic attributes, the overall formulation can be succinctly expressed as follows:

$$a_{uv}^L = \underset{v \in P(u)}{soft\max}\left(H_u^{L-1}(H_v^L)^T\right) \tag{2}$$

An augmentation method is implemented to address the significant number of connections in the DAG, which vary depending on different Emotion Recognition in Conversation datasets. In this study, we incorporate position encoding (PE) into the encoded properties of utterances. The augmentation employed in this study enhances the learning process of the RRAA function, hence ensuring its capacity to accurately identify and capture the intrinsic sequential patterns present in the input data. The chosen approach for incorporating positional information into the Transformer design is based on the theoretical foundations of the model. This involves using sine and cosine functions for encoding positional information as follows:

$$\begin{aligned} PE_{(pos,2i)} &= sin\left(pos/10000^{2i/d_{embedding}}\right) \\ PE_{(pos,2i+1)} &= cos\left(pos/10000^{2i/d_{embedding}}\right) \end{aligned} \tag{3}$$

where pos is the position index, $d_{embedding}$ is the dimension of output, $2i$ represent the index of output. In the course of

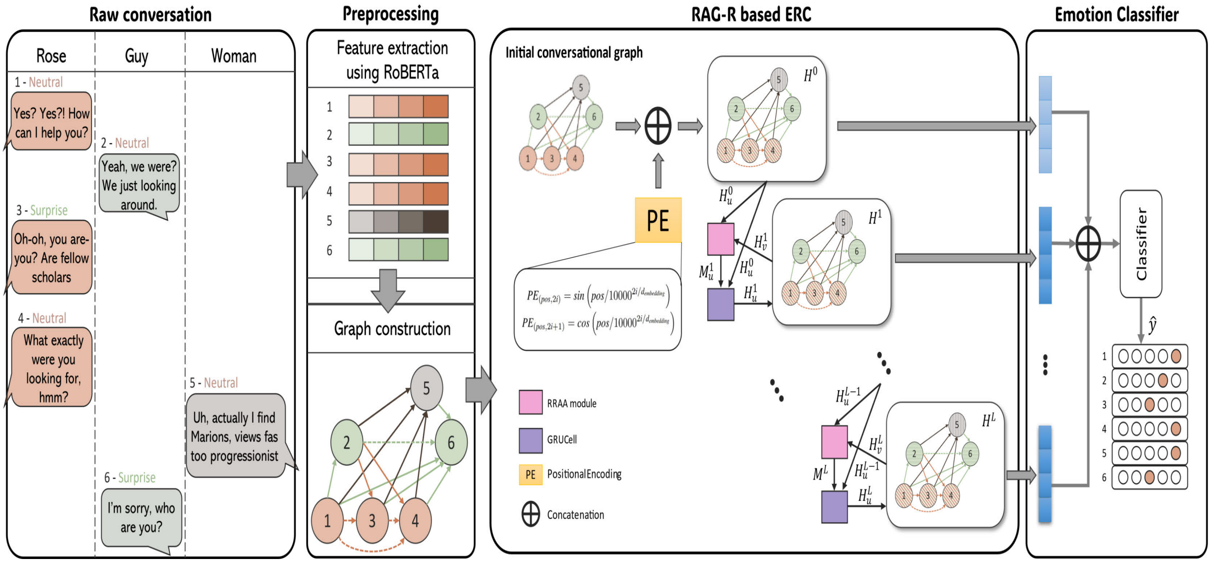

Raw conversation
Rose
Guy
Woman
1 - Neutral
Yes? Yes?! How can I help you?
2 - Neutral
Yeah, we were? We just looking around.
3 - Surprise
Oh-oh, you are-you? Are fellow scholars
4 - Neutral
What exactly were you looking for, hmm?
5 - Neutral
Uh, actually I find Marions, views fas too progressionist
6 - Surprise
I'm sorry, who are you?
Preprocessing
Feature extraction using RoBERTa
Graph construction
RAG-R based ERC
Initial conversational graph
PE
$PE_{(pos,2i)} = sin\left(pos/10000^{2i/d_{embedding}}\right)$
$PE_{(pos,2i+1)} = cos\left(pos/10000^{2i/d_{embedding}}\right)$
H^0
H^1
H^L
H_u^0
H_v^1
M_u^1
H_u^1
H_u^{L-1}
H_v^L
M_u^L
H_u^L
RRAA module
GRUCell
PositionalEncoding
Concatenation
Emotion Classifier
Classifier
$\hat{y}$

FIGURE 3. Overall Architecture of Deep Graph-Recurrent model for ERC.

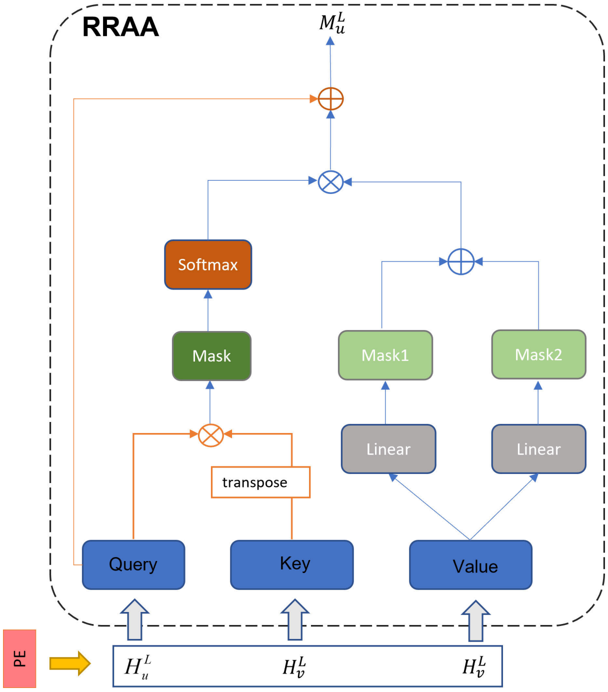


FIGURE 4. Residual Relation-aware Attention. Mask1 and Mask2 are used to identify the relation type of the edges. Nodes v are predecessor of node u.

our experimental investigations, we scrutinize the impact of index normalization on positional encoding and discern that the performance of the sine and cosine positional encoding surpasses that achieved through the utilization of index normalization of positional encoding. A graphical representation of the RRAA module is provided for elucidation in Figure 4.

3) RELATION-AWARE ATTENTION GRAPH WITH RECURRENT LAYER (RAG-R)

After performing the computation of M_u^L using the RRAA aggregate function, the subsequent phase in the process entails employing GRUs to control the information flow from previous layers to the current layer. Given the input M_u^L and the representation H_u^{L-1} of node u at the $(L-1)^{th}$ layer as the hidden state, the formulation can be written as follows:

$$H_u^L = GRUCell\left(M_u^L, H_u^{L-1}\right) \tag{4}$$

where H_u^{L-1} denotes the updated representation of node u at the L^{th} layer. Notably, RRAA aggregates information from nodes within the same layer, necessitating the sequential computation of each node's representation in the subsequent layer through a loop function. This involves iteratively computing the aggregate information M_u^L for each node u using RRAA (as per formula (1)), subsequently employing this output as the initial input for formula (4), where the GRU utilizes H_u^{L-1} at the $(L-1)^{th}$ layer as the hidden state.

To mitigate the potential over-smoothing issue [29] associated with graph-based methods, specifically the decline in performance upon stacking additional layers, we avoid direct utilization of the RRAA output. Instead, the gating mechanism within the GRUs is employed to selectively retain information from the previous Residual Relation-aware Graph-Recurrence (RAG-R) layer along with the new input data.

This architectural choice contrasts with DAGNN, where the positions of the two arguments are inverted. In DAGERC, a dual GRU approach is employed, with the positions of M_u^L and H_u^{L-1} alternated in each turn, and the outputs of both GRUs concatenated to obtain the final representation. However, for the Emotion Recognition in Conversation (ERC) task, the emotional content of an utterance should be inferred from its context. Thus, to prevent over-smoothing and to avoid direct utilization of the RRAA output, a single GRU with H_u^{L-1} as the hidden state is employed to regulate context propagation across all RAG-R layers.

Subsequently, the concatenation of all H^L layers serves as the ultimate representation for each node, passing through a fully connected layer. In the initial layer, feature dimensionality is reduced, while in the final layer, the number of output nodes matches the class labels, facilitating the prediction of the emotion associated with each utterance, as denoted by Equations 5, 6, and 7 in Shen's work [5] as follows:

$$H_u = \|_{L=0}^{L} H_u^L \tag{5}$$

$$P_u = ReLU\left(W^h H_u + b^h\right) \tag{6}$$

$$\hat{y}_t = \underset{S}{argmax}\,(P_t\,|s|) \tag{7}$$

In addition, cross-entropy loss is employed as the objective function during training to identify emotional state of each utterance.

V. EXPERIMENTAL RESULTS

This section provides a thorough description of the dataset used, explains the complexities of the training process, and includes a complete evaluation of the performance demonstrated by the proposed methodology. Following this, a comprehensive examination has been undertaken on many aspects of the model's structure, including significant factors such as positional encoding, the amount of preceding connections, and the quantity of graph layers. The purpose of this analytical examination is to identify and understand the subtle impacts of these architectural components on the overall effectiveness of the suggested approach.

A. DATASETS

In this paper, our proposed methodology undergoes rigorous evaluation across three distinct datasets: IEMOCAP [16], MELD [14], and EmoryNLP [15]. A comprehensive overview of the statistical characteristics pertaining to these datasets is elucidated in Table 1. According to Table 1, it shows that the datasets, such as IEMOCAP, MELD, and EmoryNLP, have class imbalance for different emotion categories. However, our main research is to investigate the connections between speakers in a conversation using a fully connected directed acyclic graph, rather than addressing the issue of imbalance. Therefore, we utilized categorical cross-entropy loss to align with the goals of researching on the interconnections between speakers and the dynamics of conversational emotion. Figure 5 presents the distribution and variability of dialogue lengths across the training, validation, and test sets for each dataset. The test set in the EmoryNLP dataset has a slightly higher average number of utterances per dialogue compared to both the training and validation sets. In contrast, the IEMOCAP dataset indicates an obvious increase in the average amount of utterances per dialogue when comparing the training set to the validation and test sets. The MELD dataset shows slight differences in the average number of utterances per dialogue among the three sets.

IEMOCAP: This subject-independent dataset contains prepared dialogues performed by professional actors. The first eight speakers' dialogues form the training set, while the next two contribute to the test set. We select 20 dialogues from 120 in the training set for validation. The range of emotions includes happiness, sadness, anger, excitement, frustration, and neutrality.

MELD: It extends the EmotionLines dataset [30] for emotional analysis using over 1400 dialogues and 13000 utterances from the popular TV program "Friends". Emotions are labeled: neutral, joyful, surprise, sadness, anger, disgust, and fear. The dataset is separated into training, validation, and testing sets to match conventional experimental designs.

EmoryNLP: This dataset, based on "Friends", identifies emotions neutral, sad, mad, scared powerful, peaceful, and joyful. The dataset achieves an episode independence condition with 77, 11, and 9 episodes in the training, validation, and test sets. All utterances from an episode are kept in the same set, giving the dataset its unique structure for emotion recognition studies.

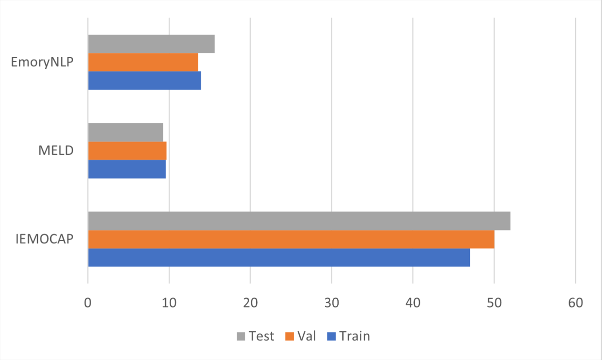


FIGURE 5. Number of utterances per dialogue on three datasets.

B. EXPERIMENTAL SETTINGS

1) CONFIGURATION

During training, two variants of positional encoding (PE) were implemented. For index normalization, the index value for each speech was determined and divided by the maximum number of utterances in a discussion. Sine and cosine functions were also utilized for PE, with a fixed value of 4 for the number of embedding dimensions. The graph-recurrent module's layer count was left flexible to achieve architecture adaptability. The Adam optimizer facilitated weight optimization, and evaluation metrics included the weighted-average F_1 score and Accuracy. Implementation was in PyTorch, and experiments involved comparisons with several baseline methods, which detailed in Section V-B2. The training process utilized an RTX 3070 for processing power. Detailed hyperparameter specifications are provided in Table 2.

2) BASELINE METHODS

We compare our proposed method with the following baseline approaches:

TABLE 1. Data statistics and emotion category distribution of IEMOCAP, MELD, and EmoryNLP datasets.

Dataset	Partition	#Dialogues	#Utterances	Emotion (#Utterances)						
				Happy	Angry	Sad	Frustrated	Excited	Neutral	
IEMOCAP	train	100	4778	392	711	739	1149	620	1167	
	val	20	980	60	222	100	319	122	157	
	test	31	1622	143	170	245	381	299	384	
				Disgust	Angry	Sad	Fear	Surprise	Neutral	Joyful
MELD	train	1038	9989	271	1109	683	268	1205	4710	1743
	val	114	1109	22	153	111	40	150	470	163
	test	280	2610	68	245	208	50	281	1256	402
				Mad	Scared	Sad	Peaceful	Powful	Neutral	Joyful
EmoryNLP	train	713	9934	1076	1285	671	900	784	3034	2184
	val	99	1344	143	178	75	132	134	393	289
	test	85	1328	113	182	98	159	145	349	282

TABLE 2. Implementation setting.

Parameter	IEMOCAP	MELD	EmoryNLP
Learning rate	0.0005	0.0005	0.0001
Graph-Recurrent Layer	3	1	3
PE Embedding	4	4	4
Batch Size	16	16	16
Epoch	40	50	100

- **DialogueGCN** [1]: Utilizes a graph neural network for modeling self and inter-speaker dependencies, enhancing emotion-relevant context through information propagation.
- **RoBERTa** and **RoBERTa+GRU**: We implement RoBERTa-based and combination of RoBERTa and GRU networks as baseline methods for ERC.
- **COSMIC** [3]: Leverages commonsense knowledge for utterance-level emotion recognition, providing a comprehensive understanding of emotional dynamics.
- **DAGERC** [5]: Introduces a directed acyclic graph to encode utterances, combining graph-based and recurrence-based models.
- **SenticGAT** [31]: Introduces a context- and sentiment-aware framework using a dynamic representation of common-sense knowledge through a graph attention mechanism.
- **MVN** [32]: Introduces a Multi-View Network to capture word- and utterance-level dependencies for emotion recognition.
- **MM-DFN** [33]: A graph-based dynamic framework capturing contextual information dynamics in different semantic spaces.
- **HSGCF** [34]: Introduces a hierarchical structure with five graph convolution layers for discriminative emotional features.
- **GraphCFC** [35]: Introduces a directed graph-based cross-modal feature complementation module for learning contextual and interactive information.
- **MVTCN** [36]: Introduces a multiview attention network to integrate dynamic interaction information and capture cross-modal dynamic dependencies.
- **AccumWR** [37]: Enhances sentence modeling by accumulating word vector representations with multilevel contextual integration.
- **DBL** [38]: Leverages conversational context, models speaker and emotion dynamics, interprets informal language and sarcasm for emotion recognition.
- **Topk-Soft** [39]: Incorporates variable-length context and two speaker-aware units for explicit modeling of inner- and inter-speaker dependencies.

C. PERFORMANCE

The results of our proposed method, as well as comparisons with state-of-the-art models on three distinct datasets (IEMOCAP, MELD, and EmoryNLP), are summarized in Table 3. For the IEMOCAP dataset, our method achieved a weighted average F_1 (WA-F_1) score of 69.10 and an accuracy of 68.72, surpassing the top-performing model, MM-DFN [33], which achieved a WA-F_1 score of 68.18 and accuracy of 68.21. For the MELD dataset, our proposed method attained a WA-F_1 score of 63.82 and an accuracy of 64.04. In comparison with the highest-performing model, AccumWR [37], which achieved a WA-F_1 score of 64.99, our model demonstrates competitive performance in effectively identifying emotions within the conversational context. On the EmoryNLP dataset, our proposed method achieved a notable WA-F_1 score of 39.85, surpassing the recent top-performing model, Topk-Soft [39], which obtained a WA-F_1 score of 38.93.

In summary, the proposed method stands out as the top-performing model across both WA-F_1 and Accuracy metrics, showcasing its effectiveness in advancing the state-of-the-art in ERC. The development of a novel benchmark for the EmoryNLP and IEMOCAP datasets is particularly remarkable. However, while examining MELD, it becomes apparent that the conversational environment is marked by a relatively small number of utterances, with an average of roughly 9 utterances each discussion. It is worth noting cases in which certain dialogues consist of only a single utterance. In contrast to the IEMOCAP dataset, which is characterized by a more extended dialogue structure consisting of roughly 50 utterances per discussion, and the EmoryNLP dataset, which exhibits an intermediate dialogue length of around 15 utterances per discourse, the present dataset demonstrates a different pattern. In the given situation, the potential benefits of our fully connected DAG may not be

TABLE 3. A comparative analysis between the proposed methodology and prevailing approaches in the realm of ERC, employing evaluation metrics such as Weighted-average F_1 (WA-F_1) and Accuracy (Acc). Emphasis is denoted through the use of bold text to underscore instances of superior performance.

Model	IEMOCAP		MELD		EmoryNLP	
	WA-F_1	Acc	WA-F_1	Acc	WA-F_1	Acc
DialogueGCN (2019) [1]	64.10	65.25	58.10	-	-	-
RoBERTa	65.23	65.17	62.91	62.57	37.55	38.93
RoBERTa+GRU	65.47	65.35	63.51	63.68	38.20	39.98
COSMIC (2020) [3]	65.21	-	64.28	-	37.10	-
DAGERC (2021) [5]	68.03	67.69	63.40	63.72	38.68	39.31
SenticGAT (2022) [31]	54.45	-	59.19	-	36.59	-
MVN (2022) [32]	65.44	65.32	59.03	61.29	-	-
MM-DFN (2022) [33]	68.18	68.21	59.46	62.49	-	-
HSGCF (2023) [34]	65.18	65.13	-	-	-	-
GraphCFC (2023) [35]	60.09	59.95	56.81	60.77	-	-
MVTCN (2023) [36]	64.1	64.7	58.6	60.6	-	-
AccumWR (2023) [37]	67.78	-	**64.99**	-	38.21	-
DBL (2023) [38]	62.53	-	64.03	-	36.99	-
Topk-Soft (2023) [39]	66.35	-	-	-	38.93	-
Proposed Method	**69.10**	**68.72**	63.82	**64.04**	**39.85**	**40.66**

TABLE 4. Performance of the proposed model on each emotional state. Evaluation metric is WA-F_1 score.

IEMOCAP	Happiness	Sadness	Anger	Excited	Frustrated	Neutral	-
	49.83	81.73	67.82	70.83	67.1	68.76	-
EmoryNLP	Joyful	Sad	Mad	Powerful	Scare	Neutral	Peaceful
	50.73	22.51	40.56	16.04	39.38	53.82	20.78
MELD	Happiness	Sadness	Anger	Surprise	Disgust	Neutral	Fear
	61.5	38.22	48.7	57.63	33.93	77.31	25.9

fully utilized, suggesting a need for additional investigation and optimization when dealing with limited conversational content.

In addition, Table 4 provides a nuanced analysis of the proposed model's performance across various emotional states, as measured by the weighted-average F_1 score, on three distinct datasets. In the IEMOCAP dataset, the proposed model excels in recognizing emotional states such as Happiness (49.83) and Sadness (81.73), showcasing its proficiency in capturing nuanced expressions of joy and sorrow. The model also demonstrates substantial competence in identifying Anger (67.82) and Excitement (70.83). However, it faces challenges in discerning Frustration (67.1) and exhibits a moderate performance in recognizing Neutral emotions (68.76). Moving to the EmoryNLP dataset, the proposed model exhibits strong performance in identifying Joyful (50.73) and Neutral (53.82) emotions. However, it faces difficulties in recognizing emotions such as Sad (22.51) and Powerful (16.04), indicating potential areas for improvement. In the MELD dataset, the model excels in recognizing Happiness (61.5), Surprise (57.63), and Neutral (77.31) emotions, while facing challenges in identifying emotions like Sadness (38.22) and Fear (25.9). This comprehensive breakdown underscores the model's capability in capturing certain emotional nuances effectively, while also highlighting specific areas where further refinement may enhance its overall performance.

D. ABLATION STUDIES

In this section, we meticulously dissect the components that contribute to the effectiveness of the proposed emotion recognition model. A series of ablation studies are conducted, systematically exploring the impact of varying Positional Encoding approaches, different numbers of predecessor's connections, and diverse configurations of the number of RAG-R Layers. Through these experiments, we aim to unravel the intricate interplay of these elements and discern their individual contributions to the overall success of the model. This detailed analysis provides valuable insights into the model's robustness, shedding light on the key factors influencing its performance in the challenging task of emotion recognition within conversational contexts.

TABLE 5. Performance on different types of positional encoding approaches using in residual relation-aware attention. evaluation metric is weighted-average F_1).

Positional Encoding	IEMOCAP	MELD	EmoryNLP
Without PE	66.29 (↓)	63.47(↓)	38.78 (↓)
Index Normalization	68.36(↓)	63.53(↓)	38.91(↓)
Sine Cosine	**69.10**	**63.82**	**39.85**

TABLE 6. Performance on different number of connections on DAG. evaluation metric is weighted-average F1.

Connection	IEMOCAP	MELD	EmoryNLP
All	**69.10**	**63.82**	**39.85**
10	67.96(↓)	63.71(↓)	38.38(↓)
5	68.1(↓)	63.68(↓)	39.31(↓)

1) PERFORMANCE ON DIFFERENT TYPES OF POSITIONAL ENCODING APPROACHES

In this study, we purposefully abstain from setting constraints on the quantity of links within the DAG. As a result, the extent of links assumes significant dimensions, which is particularly noticeable in the IEMOCAP dataset. In this dataset, a single discussion consists of around 50 utterances. It is worth mentioning that the dataset includes a conversation that holds the record for the maximum number of utterances, amounting to a total of 176. Due to the inherent limitations of Attention mechanisms in capturing sequential information, incorporating positional information becomes a crucial approach to improve the model's capability to accurately identify the specific position of an utterance within a conversation. This, in turn, aids in the effective integration of sequential data. In order to undertake an empirical investigation on the effects of several Positional Encoding (PE) methods, specifically no PE, index normalization PE, and sine cosine PE, we systematically performed trials in various circumstances. The results of these experiments are presented in Table 5.

The results collected from the study provide significant insights into the impact of physical education (PE) on enhancing the performance of models. Significantly, the model achieves its highest level of performance when sine and cosine positional encodings are utilized, but the least desirable results are found when positional encodings are not present. Moreover, it is important to notice a significant decrease in performance, specifically observed in the IEMOCAP dataset, when the feature of PE is

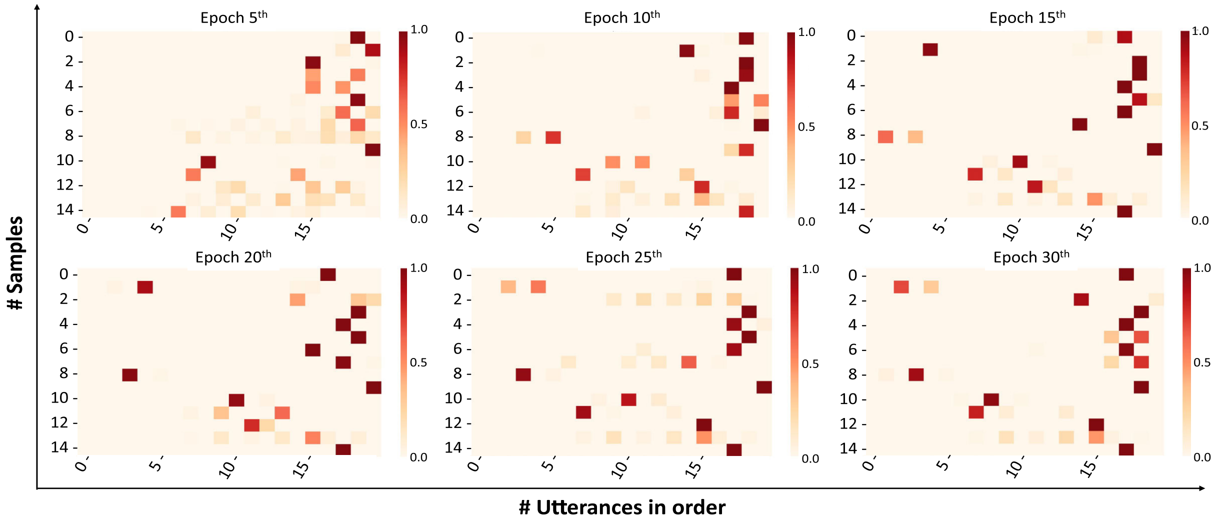


FIGURE 6. Visualization of utterance connection during training process.

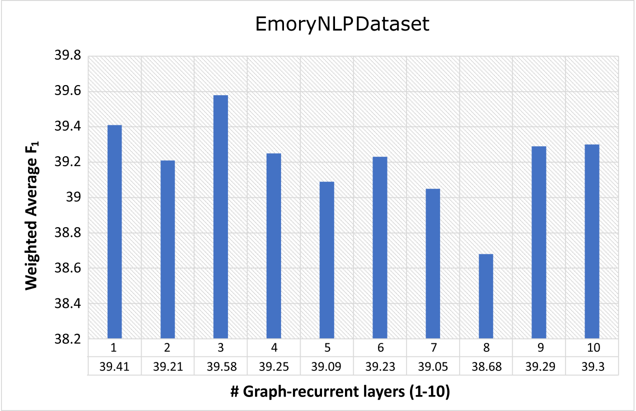


FIGURE 7. Performance on EmoryNLP dataset using different number of RAG-R layers ranging from 1-10.

excluded. The noticeable decrease in performance highlights the significant importance of including phonetic encoding, especially in datasets with a large number of utterances in a single discussion.

2) PERFORMANCE ON DIFFERENT NUMBER OF PREDECESSOR'S CONNECTIONS

In our research, we introduced a fully connected Directed Acyclic Graph (DAG) as a novel approach for modeling conversational dynamics. To assess the impact of complete connectivity within the DAG structure, we conducted a series of experiments varying the number of predecessor connections that explore configurations with 5 connections, 10 connections, and a scenario with full connectivity. The comprehensive results of these experiments are detailed in Table 6.

Analysis of the outcomes reveals that our model achieves optimal performance when implemented without restrictions on connections within the DAG. This substantiates the efficacy of our proposed methodology employing a fully connected DAG. Further scrutiny of the results indicates a more significant decline in performance on the IEMOCAP dataset compared to the MELD and EmoryNLP datasets. This discrepancy can be attributed to the substantial variance in the number of utterances per dialogue across datasets, with IEMOCAP exhibiting approximately 50 utterances per dialogue compared to the approximately 10 utterances per dialogue in MELD and EmoryNLP. The observation underscores the superior performance of our method when applied to datasets characterized by a higher number of utterances within each dialogue. Additionally, an exploration into the learning dynamics of attention weights in Equation (2) is undertaken, as shown in Figure 6. Notably, the visual representations illustrate that during the initial stages of training, the most critical information is localized in the immediate vicinity of the utterance requiring emotion prediction. Over time, this information diffuses across the entire conversation, emphasizing the adaptability and effectiveness of our proposed methodology utilizing a fully connected DAG for modeling conversational intricacies.

3) PERFORMANCE ON DIFFERENT NUMBER OF RAG-R LAYER

In these experiments, the aggregation of information from preceding nodes is achieved by sequentially stacking numerous layers. The final representation of each node is obtained by concatenating all graph-recurrent layers. Our investigation examines the intricate impact of altering the quantity of graph-recurrent layers on the overall efficacy, a meticulous examination carried out precisely utilizing the EmoryNLP dataset. In order to comprehensively analyze the effect, we methodically manipulate the number of layers within the range of 1 to 10 and display the results in Figure 7. The results demonstrate a notable level of consistency in the performance of models using various configurations of graph-recurrent layers. A significant majority of the outcomes exceed the criterion of 39. It is worth mentioning that there is a singular instance in which the performance experiences a decline, specifically when the quantity of graph-recurrent layers is established at 8. The presence of this anomaly highlights the effectiveness of utilizing Gated Recurrent Network (GRN) as a strategic approach to address the issue of over-smoothing that is inherent in graph-based approaches.

VI. CONCLUSION AND FUTURE WORKS

This paper has introduced an innovative method for Emotion Recognition in Conversation (ERC) by the utilization of a fully connected directed acyclic graph. The thorough knowledge of the subtle dynamics inside a conversation is facilitated by the employment of two types of edge relations in this graph, together with the introduction of residual relation-aware attention and positional encoding. The utilization of Gated Recurrent Units (GRUs) serves to augment the model's capacity in regulating the transmission of conversational context across many levels.

The experiments conducted on diverse ERC datasets demonstrate the significant improvements achieved by our proposed methodology over baseline models. The central contributions of our work can be summarized as follows: (1) Utilization of a fully connected directed acyclic graph with two types of edge relations to comprehensively leverage relations among speakers. (2) Introduction of residual relation-aware attention with positional encoding and GRUs to capture relation-aware features and control conversation context propagation through graph-recurrent layers. This research not only contributes to the expanding knowledge in ERC but also establishes an essential foundation for future progress in comprehending and representing emotional dynamics in conversational environments.

Although our current focus is on speaker relationships, we recognize the importance of improving model robustness through balanced dataset considerations. Therefore, in future work, we will address the class imbalance issue in emotion recognition by exploring techniques such as oversampling, undersampling, or weighted loss functions. Furthermore, future work will not only explore additional dimensions of conversation, such as topic, viewpoint, and personality but will also focus on refining and improving the proposed method. One avenue for improvement involves a more nuanced understanding of conversation topics. Investigating how the proposed model can adapt to and identify varying topics within a conversation will contribute to a more contextually aware and adaptive emotion recognition system. Incorporating topic modeling techniques and dynamic attention mechanisms could be explored to enhance the model's sensitivity to evolving discussion themes.

In addition, the enhancement of the proposed approach will entail a more comprehensive investigation of distinct perspectives inside a discourse. One might potentially explore modifications to the model's architecture in order to more effectively capture and distinguish the various emotional nuances that emerge from a range of perspectives. The potential enhancement of the model's performance in emotion recognition and interpretation can be achieved by fine-tuning the attention mechanisms to accommodate diverse speaker roles and viewpoints.

In terms of personality traits, future research will aim to tailor the proposed method to accommodate individual differences in emotional expression. Investigating how specific personality characteristics influence emotional responses and expression patterns will guide the development of a more personalized and adaptive emotion recognition framework. This may involve incorporating features or embeddings that account for speaker-specific traits, fostering a more individualized understanding of emotional cues.

REFERENCES

[1] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*.

[2] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7360–7370.

[3] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," 2020, *arXiv:2010.02795*.

[4] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," 2021, *arXiv:2106.01071*.

[5] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," 2021, *arXiv:2105.12907*.

[6] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," 2014, *arXiv:1409.1259*.

[7] C. Darwin, "The expression of the emotions in man and animals," in *The Expression of the Emotions in Man and Animals*. Chicago, IL, USA: Univ. Chicago press, 2015.

[8] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.

[9] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," 2018, *arXiv:1803.06397*.

[10] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6685–6689.

[11] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 70–74.

[12] C. Strapparava and A. Valitutti, "WordNet affect: An affective extension of WordNet," in *Proc. LREC*, vol. 4, Lisbon, Portugal, 2004, p. 40.

[13] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical Turk to create an emotion lexicon," in *Proc. NAACL HLT*, 2010, pp. 26–34.

[14] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.

[15] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in *Proc. AAAI Workshop Affect. Content Anal.*, 2018, pp. 44–51.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[17] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.

[18] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," 2019, *arXiv:1904.04446*.

[19] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6818–6825.

[20] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, "A hierarchical transformer with speaker modeling for emotion recognition in conversation," 2020, *arXiv:2012.14781*.

[21] T. Kim and P. Vossen, "EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa," 2021, *arXiv:2108.12009*.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[23] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5415–5421.

[24] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," 2020, *arXiv:2012.08695*.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5753–5763.

[26] C. Liang, C. Yang, J. Xu, J. Huang, Y. Wang, and Y. Dong, "S+PAGE: A speaker and position-aware graph neural network model for emotion recognition in conversation," 2021, *arXiv:2112.12389*.

[27] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "DAG-recurrent neural networks for scene labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3620–3629.

[28] V. Thost and J. Chen, "Directed acyclic graph neural networks," 2021, *arXiv:2101.07965*.

[29] C. Cai and Y. Wang, "A note on over-smoothing for graph neural networks," 2020, *arXiv:2006.13318*.

[30] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, *arXiv:1802.08379*.

[31] G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria, "Context- and sentiment-aware networks for emotion recognition in conversation," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 699–708, Oct. 2022.

[32] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowl.-Based Syst.*, vol. 236, Jan. 2022, Art. no. 107751.

[33] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7037–7041.

[34] B. Wang, G. Dong, Y. Zhao, R. Li, Q. Cao, K. Hu, and D. Jiang, "Hierarchically stacked graph convolution for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110285.

[35] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Trans. Multimedia*, early access, pp. 1–13, 2023.

[36] J. Wen, D. Jiang, G. Tu, C. Liu, and E. Cambria, "Dynamic interactive multiview memory network for emotion recognition in conversation," *Inf. Fusion*, vol. 91, pp. 123–133, Mar. 2023.

[37] X. Jieying, N. P. Minh, M. Blake, and N. M. Le, "Accumulating word representations in multi-level context integration for ERC task," in *Proc. 15th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2023, pp. 1–6.

[38] S. Peng, R. Zeng, H. Liu, L. Cao, G. Wang, and J. Xie, "Deep broad learning for emotion classification in textual conversations," *Tsinghua Sci. Technol.*, vol. 29, no. 2, pp. 481–491, Apr. 2024.

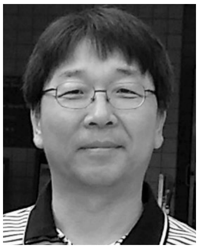
[39] M. Zhang, X. Zhou, W. Chen, and M. Zhang, "Emotion recognition in conversation from variable-length context," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

ANH-QUANG DUONG received the B.S. degree from the Department of Information Technology, VNU University of Engineering and Technology, Vietnam, in 2019, and the M.S. degree from the Department of Artificial Intelligence Convergence, Chonnam National University. His research interests include the multimodal-based emotion recognition, the Internet of Things, machine learning, and deep learning.

NGOC-HUYNH HO (Member, IEEE) received the B.S. degree from the Department of Telecommunication Engineering, Ho Chi Minh City University of Technology, Vietnam, in 2015, the M.S. degree from the School of Electronics and Computer Science, Kookmin University, South Korea, in 2017, and the Ph.D. degree from the Department of Artificial Intelligence Convergence, Chonnam National University. He is currently a Postdoctoral Researcher with Chonnam National University. His research interests include the multimodal-based emotion recognition, machine learning, deep learning and its applications, and bioinformatics.

SUDARSHAN PANT (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Mokpo National University, South Korea. He is currently a Postdoctoral Researcher with Chonnam National University. His research interests include machine learning, affective computing, and healthcare AI.

SEUNGWON KIM received the bachelor's and master's degrees in computer science from the University of Tasmania and the Ph.D. degree in human interface technology from the HITLab NZ, in November 2016. He is currently a Distinguished Postdoctoral Research Fellow specializing in remote collaboration systems and augmented reality (AR) technology. He continues to shape the field of human–computer interaction, positioning himself as a leading figure in collaborative technologies.



SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.