Enhancing Emotion Recognition in Conversations through Global Context: An Empirical Analysis

Anonymous ACL submission

Abstract

According to multimodal and contextualized nature of the human conversation, correctly identifying an emotion for given utterance in the conversation has always been a challenging task. Recent research benefits from Graph Neural Networks by capturing implicit relationship of temporally proximate utterances. In this paper, we expand the structure of the graph exploited by these models reflecting the global context of the conversation and explore how leveraging conversational context and interactions can lead to more accurate emotion recognition. We empirically analyze the modules on Emotion Recognition in Conversation models, showing this approach enhances the performance of these models. Our experiments show that incorporating global conversational context has a positive effect on the performance of emotion recognition.

1 Introduction

Emotion Recognition in Conversation (ERC) is a task of recognizing correct labels of emotion for sentences in a dialogue. Recently, ERC has become a significant area of interest for researchers due to its potential applications in fields requiring multimodal interaction (Poria et al., 2019), and natural interactions between humans and computers. It can be used in robotics, can be applied in medical science (Zucco et al., 2018), and household devices capable of generating responses that demonstrate emotional intelligence and empathy. This necessitates the precise interpretation of the embedded meanings within each sentence, speech, video and more, thereby significantly elevating the importance of the field of emotion recognition.

However, conversation represent complex interplay of multiple elements including hand gestures, facial expressions, language, speech, sound, context, and emotions, making the prediction of emotions within dialogue sentences a challenging endeavor. Many researchers tried various attempts to enhance the performance of emotion recognition by leveraging a variety of factors. Also, they tried implementing techniques in machine learning to increase the performance of emotion recognition models. Among these attempts, Graph Neural Network (GNN) is one of schemes turned out to be successful in improving the performance of the task (Joshi et al., 2022; Hu et al., 2022; Chen et al., 2023). Learning embedding of both nodes and their relationships, ERC models using graph network architecture proved their capacity to capture the relationship between sentences and predict underlying emotional feature.

Nevertheless, these models still struggle to adequately capture the relationships between all utterances or modalities, often limited by factors such as graph size, shortage of data and etc. There have been various attempts to approach this problem from multiple perspectives and adopt different solutions. As one of these approaches, from a psychological perspective, it is anticipated that more accurate emotion recognition could be achieved by integrating into the model the notion that global contexts, such as mood, influence emotional bias, as posited by Schmid and Schmid Mast (2010). We try to bring this perspective to be implemented in graph formation. Our research investigates whether more precise Emotion Recognition in Conversation (ERC) can be achieved through simple modifications to the GNN, by more actively utilizing global context during the graph formation stage. We created global node in the graph formation stage to better capture the overall context of the conversation and explore the impact of slight changes in edge connections between nodes. Additionally, we investigate the effects of incorporating global embeddings in the classifier stage of the model. We apply these implementations to several existing GNN-based ERC models and conduct additional experiments to determine the actual differences each implementation makes.

We make the following contributions in this paper:

- Our model enhances the efficacy of ERC models by deploying a simplified yet effective methodology, which involves the strategic addition of a limited set of nodes and edges to the existing graph structure.
- We discover the mechanism by which global embeddings and global nodal interactions affect the entire graph structure.

2 Related Works

2.1 Emotion Labeled Datasets

Several publicly available datasets can be utilized in the ERC task. The IEMOCAP dataset (Busso et al., 2008) is widely recognized in the field of emotion recognition, containing multimodal data (acoustic, textual, and visual). EmotionLines (Hsu et al., 2018) comprises dialogue of text data from the popular TV show "Friends". MELD (Poria et al., 2019) is an expanded version of the EmotionLines dataset, that includes additional visual and acoustic data. The SEMAINE dataset (McKeown et al., 2011) is offered with multimodal data with dimensional emotion labels (valence, arousal, expectancy, and power), annotated with values ranging from -1 to 1 (Buechel and Hahn, 2017). Additional datasets such as EmoryNLP (Zahiri and Choi, 2018), DailyDialog (Li et al., 2017), and CMU-MOSEI (Zadeh et al., 2018) emphasize dimensional emotion labels. More recent datasets includes K-Emocon (Park et al., 2020) and AV-CAffe (Sarkar et al., 2023). We employ the IEMOCAP and MELD datasets in our analysis due to their applicability in the baseline models we use, multimodal nature and the availability of discrete emotional labels corresponding to individual utterances.

2.2 GNN-based ERC Models

The challenge in Emotion Recognition in Conversation (ERC) stems from the complexity of discerning how specific utterances within a dialogue influence the emotional state of the speaker. Early research attempted to extract context from conversations using Deep Belief Network (DBN) and Long Short Term Memory (LSTM) as demonstrated by Lee et al. (2009) and Wöllmer et al. (2010), respectively.

Later on, Graph Neural Network(GNN) were found to be affective in conveying the global state. DialogueGCN (Ghosal et al., 2019) employs GNN structures to effectively combine contexts inherent in sentences. Zhang et al. (2019) utilizes graphs to model multi-speaker scenarios. Shen et al. (2021) merged the capabilities of traditional GNN with recurrent neural models to enhance the performance. Hu et al. (2021) leverages a graph-based fusion technique to capture both intra- and inter-modality contextual features. MM-DFN (Hu et al., 2022), an evolution of MMGCN, incorporates a dynamic fusion network for more sophisticated multimodal integration. Fu et al. (2022) utilized a Graph Convolutional Network (GCN) with knowledge graphs, and Joshi et al. (2022) aims to capture both local and global information. More recently, Chen et al. (2023) focus on capturing more comprehensive multivariate relationships and utilizing multi-frequency information within the graph. Nevertheless, these models still have difficulty leveraging the full potential of the global contexts lying in the dialogue. We review methods from studies in other domains (Wang et al., 2020; LIU et al.; Wu et al., 2021) that utilized GNN structures to more effectively capture global and local information, exploring how to better incorporate global context. Additionally, we apply the use of random edges(Zhao et al., 2021) similarly to see the effect in the ERC model. While most existing studies have relied on training to achieve graph formation, we aimed to determine if simple structural changes could also result in performance differences.

3 Method

We propose methods for extracting global context from inputs that are typically common to the stages of graph-based ERC models, specifically focusing on the GNN and classifier stages.

We represent a conversation $U = \mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{T}$, consisting of feature vectors of utterances $\mathbf{u}_i$, where T is the number of utterances. The vector can contain acoustic, textual, and visual features, denoted as $\mathbf{u}_i^a \in U^a, \mathbf{u}_i^t \in U^t$, and $\mathbf{u}_i^v \in U^v$, respectively. Each utterance's feature vectors are derived from their respective feature extractor models, which do not need to be specified. Additionally, each utterance is delivered by the speaker $\phi(i) = s_n \in S = \{s_1, \ldots, s_\mathrm{N}\}$, where N denotes the number of speakers in U, and ϕ denotes a mapping from an utterance to its corresponding speaker. The features of individual modalities do not need to be context-aware. Each modality feature extractor

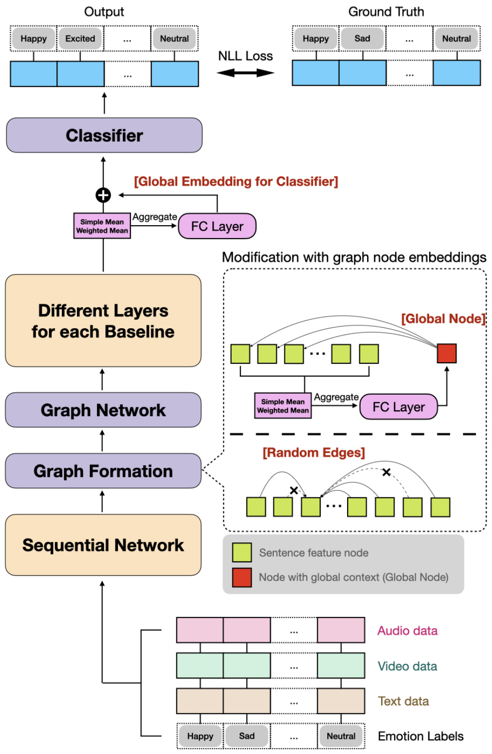

Figure 1: Overview of our modules within the general form of GNN-based ERC model. It illustrates briefly how the global context is extracted and used in global node or in global embedding classifier. It also illustriates how the random edge is formed during the graph formation phase.

independently computes $\mathbf{u}_i$.

3.1 Context Extractor

Prior layers precede the GNN layer and follow baseline architectures (Joshi et al., 2022; Hu et al., 2022, 2021). These layers serve as context extractors and can consist of any type of neural network specialized for sequential data, including LSTMs and transformers. Our model generates context vectors $\mathbf{c}_i = \text{ContextExtractor}(\mathbf{u}_i)$ from utterance embeddings $\mathbf{u}_i$.

3.2 Graph Neural Network(GNN)

3.2.1 Local Nodes and Edges

Suppose a ContextExtractor implicitly learns the relationships between different embeddings in the case of a graph neural network. In that case, it takes the different embeddings and their explicit relationships through edges as input, learning embeddings for the relationships themselves. Most ERC models (Joshi et al., 2022; Hu et al., 2022) employ the Relational Graph Convolutional Network (Schlichtkrull et al., 2018), which defines a relation $r \in \mathcal{R}$ as illustrated in the Equation (1).

$$\begin{aligned} r_{\text{past\_inter}} &= \{\mathbf{c}_i \rightarrow \mathbf{c}_j | i < j, \phi(i) \neq \phi(j)\} \\ r_{\text{future\_inter}} &= \{\mathbf{c}_i \rightarrow \mathbf{c}_j | i > j, \phi(i) \neq \phi(j)\} \\ r_{\text{past\_intra}} &= \{\mathbf{c}_i \rightarrow \mathbf{c}_j | i < j, \phi(i) = \phi(j)\} \\ r_{\text{future\_intra}} &= \{\mathbf{c}_i \rightarrow \mathbf{c}_j | i > j, \phi(i) = \phi(j)\} \end{aligned} \tag{1}$$

Neighborhood $\mathcal{N}_r(i)$ is a set of neighboring indices for $\mathbf{c}_i$ under r. The network convolves the context vectors and relations to yield new embedding reflecting the graph information (Schlichtkrull et al., 2018), where Θ_{root} and Θ_r are learnable parameters of the model. Equation (2) indicates the output of the GNN $\mathbf{z}_i$.

$$\mathbf{z}_i = \Theta_{\text{root}} \cdot \mathbf{c}_i + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \Theta_r \cdot \mathbf{c}_j \tag{2}$$

3.2.2 Global Nodes Aggregating Utterances

We aim to integrate a broader context by introducing a novel relationship, represented by $\mathbf{\Theta}_{\text{global}}$. This involves adding directed edges from a universal context node $\mathbf{c}_{\text{g}}$, to every other nodes in the network. The global node $\mathbf{c}_{\text{g}}$ is extracted using two primary methods: aggregating the input nodes of the graph through calculations such as the simple mean, the weighted mean, or through an embedding obtained via a fully-connected layer. We then integrate this global node into the graph's vertex set and connect it to other nodes through directed edges in the edge list. We define $\mathbf{\Theta}_{\text{global}}$ as the learnable parameters associated with the global embedding. Consequently, the resulting output embedding follows the configuration specified in the Equation (3).

$$\mathbf{z}_i = \Theta_{\text{root}} \cdot \mathbf{c}_i + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \Theta_r \cdot \mathbf{c}_j + \Theta_{\text{global}} \cdot \mathbf{c}_{\text{g}} \tag{3}$$

3.2.3 Random Edges

The second approach entails generating random long-distance edges within the graph. Typically, when conversation data is represented graphically, connections are established between temporally proximate conversations to facilitate the exchange of local information. Let $\mathcal{G}(i)$ represent the set of indices k from a random subset of U that satisfies

$|k - i| \geq \delta$, where δ is a minimum length edge length in the graph. Consequently, the resulting output embedding is structured as the Equation (4).

$$\mathbf{z}_i = \mathbf{\Theta}_{\text{root}} \cdot \mathbf{c}_i + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \Theta_r \cdot \mathbf{c}_j + \sum_{k \in \mathcal{G}(i)} \Theta_{global} \cdot \mathbf{c}_k \quad (4)$$

Some models (Joshi et al., 2022) have an additional Transformer layer (Shi et al., 2020) that benefits from graph neural architecture. It is placed between the previous Graph Convolutional Network layer and the classifier layer, which can be simply denoted as $\mathbf{z}'_{1..\text{T}} = \text{GraphTransformer}(\mathbf{z}_{1..\text{T}})$.

3.3 Global Embeddings for Classifier

The last module extracts the global context from the embeddings produced by the GNN. To minimize complexity, we put the classifier input $\mathbf{z}_i$ to a separate FC layer and averaged the output across the time dimension. This averaged vector is considered as "Global Embedding". We append this to $\mathbf{z}_i$. This is denoted as the Equation (5).

$$\hat{\mathbf{y}}_{\mathbf{I}} = \text{argmax}(\sigma(\mathbf{W}[\mathbf{z}_I : \mathbf{z}_g] + \mathbf{b})) \quad (5)$$

Equation (6) represents $\mathbf{z}_g$, the global embedding concatenated to the original classifier input.

$$z_g = \sum_{i=1}^{\text{T}} \frac{\mathbf{W}_g \mathbf{z}_i + \mathbf{b}_g}{\text{T}} \quad (6)$$

4 Experiment

4.1 Datasets

We use IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019) datasets. IEMOCAP dataset is a multimodal dataset assembled by recording scripted plays and improvisations, including text, speech, and facial expressions captured using motion capture devices. We used six emotion labels (happy, sad, neutral, angry, excited, and frustrated) to train and evaluate each model. MELD is a multimodal dataset derived from dialogues of the famous TV show "Friends." Its labels are annotated with seven emotions (anger, disgust, fear, joy, neutral, sadness, and surprise) and three sentiments (positive, negative, and neutral). We used only the emotion labels of the dataset.

We primarily adher to the methodologies outlined in the open-source repositories of the baseline models[1][2][3] for dataset processing. However, due to the lack of detailed instructions in some of these sources, the evaluation results reported in these works may not be entirely accurate.

[1] https://github.com/hujingwen6666/MMGCN.git
[2] https://github.com/zerohd4869/MM-DFN.git
[3] https://github.com/Exploration-Lab/COGMEN.git

4.2 Baseline Models

We applied our modules to three contemporary ERC models: MMGCN (Hu et al., 2021), MM-DFN (Hu et al., 2022), and COGMEN(Joshi et al., 2022). These models are selected based on their recency and high-ranking evaluation scores relative to other GNN-based ERC models. Each of these models incorporates trainable weights to effectively capture the relational dynamics between different utterances.

4.3 Implementation Details

We implement each model in Section 3 and compare their evaluation results with baseline vanilla models. Performance is measured by both accuracy and F1 score. We maintain consistent hyperparameters across all implementations of each baseline model.

In the Global Node module, we add an additional global node at the end of each conversation's context vector. Global node averages the context embeddings from the context extractor for each conversation.

In the Global Embedding module, the classifier processes the output of the GNN through an additional FC layer, averaging it across the time dimension, and appends it to the initial classifier input.

In the Random Edges module, we specify the number of random edges and connect nodes that lie outside a predefined window. For instance, let us suppose we set the predefined window as three. The newly formed edges could include nodes situated more than three positions away from the central node, both preceding and following it. Utilizing random edges, two nodes are randomly selected among the conversations, and if the corresponding edge is not present in the graph already, it is added to the graph. The total number of newly created edges is 10% of the number of edges in the existing graph. In cases where the conversation is shorter than the window size, all edges are connected.

Subsequently, we examine the differences in implementations according to changes in modality, as well as the variations in scores for each emotion label. To determine the practical impact of our modules, we conduct an additional experiment using the random edge module and analyze the label

	Module	MMGCN		MM-DFN		COGMEN	
		F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑
IEMOCAP	(vanilla)	65.56	65.87	66.42	66.85	69.80	70.40
	+ G.E.	66.25	66.42	**67.34**	**67.9**	69.12	**72.14**
	+ G.N.	65.6	65.99	66.14	66.54	66.68	70.18
	+ R.E.	63.88	64.88	66.24	67.16	68.29	71.54
	+ G.N. & G.E.	**66.62**	**67.22**	**67.55**	**68.15**	**70.35**	71.53
	+ R.E. & G.E.	**66.64**	**67.22**	66.62	67.34	**72.35**	**73.04**
MELD	(vanilla)	57.38	59.92	57.59	61.3	51.88	55.44
	+ G.E.	57.29	**60.61**	57.86	**61.15**	51.24	54.3
	+ G.N.	57.11	60.04	58.01	61.3	**52.00**	55.37
	+ R.E.	56.35	59.31	**58.43**	61.11	51.65	**55.86**
	+ G.N. & G.E.	**57.69**	60.5	**58.35**	**61.99**	51.33	54.66
	+ R.E. & G.E.	**57.40**	**60.57**	57.95	61.03	**51.97**	**55.65**

Table 1: F1 score (F1) and accuracy (Acc) presented in percentages (%) for implementation of our modules. G.E. is global embedding for the classifier in Section 3.2. G.N.is the module in 3.1 and R.E. means random edge module.

predictions from the actual evaluation results.

We use 4 Nvidia GeForce 1080Ti GPUs, and the models (MMGCN, MM-DFN, COGMEN) takes up to a day to train for each module implementation.

5 Results

5.1 Global Node Global Embedding Module

This section presents the results from the modules described in Section 3.2.2.

As shown in Table 1, for MMGCN (Hu et al., 2021), the model fairly showed a slight increase in performance when applied both Global Node module and Global Embedding modules, regardless of the choice of the dataset (IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2019)). In MM-DFN (Hu et al., 2022) implementations, the performance of the model with both Global Node and Global Embedding was slightly higher, and the model with only the Global Embedding also showed higher scores. Similarly, in COGMEN (Joshi et al., 2022), trained with IEMOCAP, Global Embedding implemented model showed high accuracy, while the model with both Global Node and Global Embedding model achieved a relatively high F1 score.

All three models tend to perform better with both Global Node and Global Embedding module. For Global Node, when the module was used by itself, the performance improvement was minimal. However, when combined with the global embedding, it appears to be effective.

Table 2 shows the F1 score results for each discrete emotion label. Overall, as reflected in Table 1, both the model incorporating the Global Node module and Global Embedding module and the one with only the Global Embedding module showed enhanced performance. Notably, the model combining the Global Node module and Global Embedding module outperformed the vanilla model by over 10% in F1 score for the 'happy' label. In the original evaluation, the baseline model often misclassified 'sad' instances as 'happy.' In contrast, our model exhibited fewer such errors. This improvement indicates that incorporating the global context can reduce false positives for utterances labeled 'happy.'

5.2 Random Edge Module

When both Random Edge module and Global Embedding module were applied together to MMGCN and COGMEN, they showed an improvement in F1 score performance as shown in Table 1. COGMEN demonstrated a more significant improvement. In implementing the modules to MM-DFN, it showed just a slight increase than the baseline model. This might be attributed to the possibility that MM-DFN passes through more layers in the Graph Neural Network than MMGCN and COGMEN, which could lead to a lesser degree of global context being reflected.

		Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	**Total**	
	Module	F1 ↑	F1 ↑	F1 ↑	F1 ↑	F1 ↑	F1 ↑	F1 ↑	Acc ↑
IEMOCAP	(vanilla)	26.82	82.33	69.60	63.85	**75.07**	64.74	69.80	70.40
	+ G.E.	25.09	82.19	**70.36**	59.24	74.67	**65.62**	69.12	**72.14**
	+ G.N.	34.11	84.33	69.80	**67.99**	71.70	63.90	66.68	70.18
	+ R.E.	31.02	84.14	69.28	64.59	71.97	63.59	68.29	71.54
	+ G.N. & G.E.	**37.38**	**85.17**	68.86	66.86	70.22	63.43	**70.35**	71.53
	+ R.E. & G.E.	32.27	83.58	69.06	64.85	71.79	64.86	**72.35**	**73.04**

Table 2: F1 score (F1) presented in percentages(%) for each emotions. This shows the results of the COGMEN baseline model and the models with our modules implemented in COGMEN.

		COGMEN	
	Window size / type	F1 ↑	Acc ↑
IEMOCAP	6	69.80	70.40
	10	65.12	67.44
	15	65.60	68.10
	20	66.79	70.59
	25	68.02	69.72
	2N	66.68	69.02
	3N	67.59	69.82
	4N	66.59	67.70
	5N	67.70	70.72
	Random Edge	**72.35**	**73.04**

Table 3: F1 score (F1) and accuracy (Acc) of different variations of window size or window type presented in percentages(%). It was tested on COGMEN baseline model, trained by IEMOCAP dataset.

5.3 Impact of Random Edge Module

In our exploration of the Random Edge module, we sought to demonstrate that the primary performance driver for our model is the element of randomness rather than simply increasing the number of connections through larger window sizes. To substantiate this claim, we performed two experiments. First, we tested the impact of varying window sizes by exceeding the original configuration. As shown in Table 3, although the F1 score increased with larger window sizes, there was no significant correlation between the accuracy and the larger window sizes. Second, we maintained a constant window size while increasing the hop size to observe how the random hops undertaken by the Random Edge module affect the model performance. While accuracy improved slightly, there was no significant correlation between increasing the hop size and the F1 score change. These findings suggest that variations in window size or hop size minimally impact model performance, which does not match the enhancement seen with the model employing the Random Edge module.

5.4 Comparison under different modality settings

We hypothesize that the modality of speech could significantly impact the assigned emotion label. For example, one might be less responsive to textual modalities yet more sensitive to auditory modalities.

Table 4 shows impact of modalities on the performance of the implemented modules.

We have found that models generally exhibit improved performance with the addition of textual modality. Additionally, although our model does not always achieve the best outcomes across all modalities, it is important to highlight that our combination of audio and video modalities outperforms both the vanilla model and the original model's textual and video combination. These findings support that the improvements likely stem from our model's effective use of context, which previous models may not have fully exploited.

Since COGMEN is originally trained using the IEMOCAP and CMU-MOSEI datasets (Zadeh et al., 2018), we adapted its training for the MELD dataset to ensure fair comparisons with other baseline models. The potential mismatch between COGMEN's training and the MELD dataset may partly account for the outcomes presented in Table 1; although the model incorporating both the

		A		*T*		*V*		*A+T*		*A+V*		*T+V*		*A+T+V*	
	Module	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑
IEMOCAP	(vanilla)	57.73	59.55	56.74	58.92	**45.32**	**49.73**	65.44	65.00	62.66	64.88	**62.40**	**64.73**	69.80	70.40
	+ G.E.	57.75	59.58	56.49	58.49	44.82	49.34	66.08	66.30	63.33	65.15	61.64	64.19	69.12	72.14
	+ G.N.	57.68	59.49	**57.73**	**59.82**	45.04	49.55	**66.19**	**66.42**	62.30	64.49	61.80	64.25	66.68	70.18
	+ R.E.	57.77	**59.61**	56.69	59.01	44.80	49.34	65.92	66.08	62.87	64.88	61.93	64.25	68.29	71.54
	+ G.N. & G.E.	57.72	59.55	56.73	59.04	45.00	49.49	65.47	65.09	62.79	64.88	61.54	64.01	70.35	71.53
	+ R.E. & G.E.	**57.79**	**59.61**	56.64	58.95	45.12	49.61	65.46	65.03	**63.74**	**65.54**	62.07	64.49	**72.35**	**73.04**
MELD	(vanilla)	27.21	39.45	51.43	54.93	27.42	38.08	**52.54**	**56.03**	26.45	42.69	**52.56**	**56.97**	51.88	55.44
	+ G.E.	26.37	**43.5**	51.33	55.18	27.96	**43.24**	48.94	54.03	27.19	43.03	50.89	55.14	51.24	54.3
	+ G.N.	**27.46**	39.87	50.04	53.48	27.49	38.72	51.57	54.84	26.93	42.94	51.71	56.25	**52.00**	55.37
	+ R.E.	**27.46**	39.87	49.75	52.96	28.1	40.3	51.46	54.97	**27.69**	41.83	52.54	56.5	51.65	**55.86**
	+ G.N. & G.E.	26.9	35.18	49.83	54.16	**28.47**	42.73	48.00	52.03	26.93	42.17	49.61	54.58	51.33	54.66
	+ R.E. & G.E.	26.37	**43.5**	**52.17**	**55.86**	27.3	42.77	49.69	54.54	27.11	**43.58**	50.98	55.57	51.97	55.65

Table 4: F1 score (F1) and accuracy (Acc) presented in percentages (%) for COGMEN and implementations of our modules. A is audio, T is text, V is video modality.

Random Edge and Global Embedding modules achieved relatively higher scores, its overall performance was still lower than that of the other two baseline models.

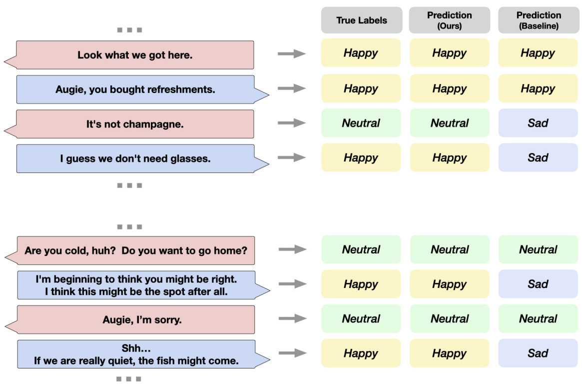

Figure 2: Segments of dialogue examples from the IEMOCAP test data. "Baseline" is COGMEN vanilla model and "Ours" is a COGMEN model implemented with both Random Edge and Global Embedding module.

5.5 Does these modules really attend to global context?

To ascertain whether our model reflects a proper representation of the global context, we extracted sample sentences that our model classified differently from the baseline model. We wanted to see whether our modules really saw the global contexts, thereby predicting the true emotion label of sentences that could be seen differently on the local level. We specifically looked at the actual evaluation results of test dataset on the COGMEN vanilla model and model that implemented our Random Edge module with global embedding classifier module to COGMEN. In cases where the original model made incorrect predictions but the model incorporating global context made correct predictions, it was often found that the utterances were temporally adjacent. For instance, the vanilla model incorrectly predicted the emotions of certain utterances in the conversations shown in Figure 2, mislabeling some sentences as 'sad.' Although the individual sentences carry negative nuances, the relevant part of the conversation involves two people making up. Therefore, an accurate prediction of the emotional label would require understanding the overall (global) context of the conversation.

Moreover, the consistent misclassification of consecutive utterances by the baseline model may be attributed to the influence of a single node's error on its neighboring nodes when calculating emotions based on local context. Our model's ability to correctly classify these instances may be due to its robustness against local errors, thereby retrieving accurate emotion labels from the given example. This hypothesis is further supported by the difference in the overall recognition scores between the baseline model and the Random Edge with Global Embedding Classifier model, as shown in Figure 1.

6 Conclusion

In this study, we enhance the architecture of graph neural networks within emotion recognition in conversation models by incorporating the global context of conversations. We investigate effects of integrating conversational context and interactions on improving the accuracy of emotion recognition in conversation models. We empirically show that these modifications could have positive impact on the performance. We also analyze the modifications in various perspective to see if these modules truly convey the global contexts to enhance the performance of the prediction task. Various experiments and their results suggest our methods are capable of leveraging global context to different types of graph networks.

7 Limitation and Future Works

One significant limitation of this research arises from the variability in the formats used by different models for embedding generation, such as pickle files and the handling of ambiguous labels. This diversity complicates the achievement of a perfectly fair comparison among models.

Another key issue is the class imbalance present in datasets, where certain emotional labels are disproportionately represented. This imbalance may impact model performance, as suggested by the possibility of including a histogram figure of dataset labels to illustrate this point. Additionally, the size of datasets represents a constraint. Given the inherently difficult nature of collecting such data, the available datasets are not large. This limitation is evidenced by the variance in results dependent on minor model settings like hyperparameters, further highlighting the challenge of dataset size and consistency. In light of these limitations, there is a desire to overcome the narrow representation space confined to categorically labeled data. One proposed solution involves mapping dimensionally labeled data to categorical labels through a form of interpolation, thereby expanding the dataset size and potentially enhancing model performance. Also, our approach faces limitations in its applicability, particularly its restriction to GNN-based models. Since many Emotion Recognition models do not utilize GNNs, it is essential to consider methods that can generally improve model performance across a broader range of models. The issue of reproducibility also presents a limitation, with many studies not fully disclosing their datasets and codes. This lack of openness has hindered the application and accurate reproduction of existing methods. Finally, it is important to acknowledge that the proposed modules may not uniformly enhance performance across all baselines. This variability underscores the need for further research to develop more universally applicable strategies that can address the model and dataset-specific challenges inherent in Emotion Recognition.

8 Ethics Statement

This study was conducted with careful emphasis on ethical considerations. All data used in this research were obtained from publicly open sources. We obtained necessary permissions from owners for data usage where required. We conducted thorough evaluations to assess the fairness and robustness of our models. During the writing, AI assistant is used for checking grammar.

References

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

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.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2022. Context-and knowledge-aware graph convolutional network for multimodal emotion recognition. *IEEE MultiMedia*, 29(3):91–100.

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.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164.

Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan. 2009. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Tenth Annual Conference of the International Speech Communication Association*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zemin LIU, Yuan FANG, Chenghao LIU, and Steven CH HOI. Node-wise localization of graph neural networks.(2021). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pages 1520–1526.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Pritam Sarkar, Aaron Posen, and Ali Etemad. 2023. Avcaffe: A large scale audio-visual dataset of cognitive load and affect for remote work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 76–85.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Petra Claudia Schmid and Marianne Schmid Mast. 2010. Mood effects on emotion recognition. *Motivation and Emotion*, 34:288–292.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2020. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*.

Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 169–178.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling.

Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. 2021. Representing long-range context for graph neural networks with global attention. *Advances in Neural Information Processing Systems*, 34:13266–13279.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

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. Macao.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. 2021. Data augmentation for graph neural networks. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 11015–11023.

Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747. IEEE.

Appendix

A Hyperparameter Setting

We mostly tried to follow the baseline models' (Hu et al., 2021, 2022; Joshi et al., 2022) original settings to emphasize fair comparison with our implementations. Settings are described in Table 5 for MMGCN, Table 6 for MM-DFN and Table 7 for COGMEN.

Dataset	*GCN Layers*	*Dropout*	*Gamma*	*Learning Rate*	*L2*
IEMOCAP	4	0.4	0.7	$3e-4$	$3e-5$
MELD	4	0.4	0.7	$3e-4$	$3e-5$

Table 5: Hyperparameter values for MMGCN.

Dataset	*GCN Layers*	*Dropout*	*Gamma*	*Learning Rate*	*L2*
IEMOCAP	16	0.4	1.0	$1e-4$	$1e-4$
MELD	32	0.2	1.0	$5e-4$	$1e-4$

Table 6: Hyperparameter values for MM-DFN.

Dataset	*Dropout*	*Learning Rate*	*Weight Decay*
IEMOCAP	0.1	$1e-4$	$1e-8$
MELD	0.1	$1e-4$	$1e-8$

Table 7: Hyperparameter values for COGMEN.

B Additional Experiment Results

Experiment results of MMGCN (Hu et al., 2021) and MM-DFN (Hu et al., 2022) under different modality settings.

C Evaluation Metrics

F1 score: F1 score is the harmonic mean of precision and recall, which can be used for imbalanced datasets like IEMOCAP or MELD. F1 score ranges from 0 to 1, with 1 being the perfect precision and recall value. Formula for the score is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: Accuracy is defined as the percentage of correct prediction of labels in the evaluation process of each model.

		A		T		V		A+T		A+V		T+V		A+T+V	
	Module	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑
IEMOCAP	(vanilla)	46.57	48.68	61.99	62.42	27.70	32.29	65.32	65.50	49.76	50.96	63.36	63.96	65.56	65.87
	+ G.E.	45.98	48.80	61.53	62.05	34.40	37.03	65.70	66.24	50.51	52.06	63.54	63.83	66.25	66.42
	+ G.N.	46.78	48.68	62.73	63.09	28.89	30.75	65.59	65.99	51.3	52.56	63.17	64.08	65.60	65.99
	+ R.E.	47.33	50.03	60.30	61.92	29.54	38.51	59.09	61.37	49.16	51.14	60.73	61.55	63.88	64.88
	+ G.N. & G.E.	44.61	47.20	62.35	63.09	30.88	33.09	65.48	65.93	50.21	51.51	63.18	63.71	66.62	67.22
	+ R.E. & G.E.	44.38	46.95	62.30	63.09	30.76	33.27	65.50	65.99	50.18	51.39	63.11	63.65	66.64	67.22
MELD	(vanilla)	40.47	49.08	56.81	59.81	31.64	48.24	57.54	60.27	41.81	50.19	57.27	60.57	57.38	59.92
	+ G.E.	41.01	49.73	56.05	59.85	37.39	48.43	57.32	60.11	44.15	50.77	57.50	60.08	57.29	60.61
	+ G.N.	40.28	49.16	57.11	59.46	34.42	48.39	56.98	60.23	43.35	50.15	57.43	60.69	57.11	60.04
	+ R. E.	36.92	48.31	52.77	57.05	33.52	48.12	55.35	59.50	41.60	48.77	55.32	58.58	56.35	59.31
	+ G.N. & G.E.	40.54	49.08	57.32	60.08	32.27	48.31	57.27	60.34	43.66	50.46	57.55	60.54	57.69	60.50
	+ R.E. & G.E.	41.47	49.35	56.52	59.62	36.54	48.35	57.24	60.46	44.23	50.54	57.82	60.69	57.40	60.57

Table 8: F1 score(F1) and accuracy(Acc) presented in percentages(%) for MMGCN and implementations of our modules. A is audio, T is text, V is video modality.

		A		T		V		A+T		A+V		T+V		A+T+V	
	Module	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑
IEMOCAP	(vanilla)	51.84	54.65	60.80	61.00	26.64	31.61	64.46	64.76	52.62	54.47	62.05	62.42	66.42	66.85
	+ G.E.	54.97	55.95	62.33	62.23	26.73	31.85	64.61	64.63	53.72	55.51	61.43	61.55	67.34	67.90
	+ G.N.	54.81	55.95	61.20	61.37	30.86	32.72	62.83	62.91	50.82	52.68	62.42	62.85	66.14	66.54
	+ R. E.	50.47	52.50	60.63	60.51	32.70	35.37	61.20	61.55	49.59	52.56	62.77	62.60	66.24	67.16
	+ G.N. & G.E.	57.16	57.79	62.25	62.11	26.94	32.90	63.70	63.89	51.62	53.97	61.45	61.68	67.55	68.15
	+ R.E. & G.E.	52.10	54.84	61.87	62.17	26.98	29.57	66.09	66.79	53.73	55.76	62.31	63.03	66.62	67.34
MELD	(vanilla)	42.01	47.85	56.96	60.08	35.42	45.59	57.68	60.54	43.18	48.85	58.46	61.26	57.59	61.30
	+ G.E.	42.02	48.16	57.30	60.46	34.08	48.54	57.31	60.08	44.77	50.19	57.77	61.46	57.86	61.15
	+ G.N.	43.56	48.85	57.16	60.27	35.05	48.35	57.92	60.80	43.90	49.85	57.78	60.88	58.01	61.30
	+ R. E.	41.03	47.74	57.24	59.81	34.70	42.34	57.63	60.11	44.28	50.11	57.76	61.26	58.43	61.11
	+ G.N. & G.E.	41.53	47.36	56.76	59.20	35.25	48.12	57.52	60.42	44.27	50.08	58.07	60.77	58.35	61.99
	+ R.E. & G.E.	43.63	47.36	56.97	60.54	35.09	46.70	57.81	60.57	44.15	49.81	57.70	60.23	57.95	61.03

Table 9: F1 score(F1) and accuracy(Acc) presented in percentages(%) for MM-DFN and implementations of our modules. A is audio, T is text, V is video modality.