COGMEN: COntextualized GNN based Multimodal Emotion recognitioN

Abhinav Joshi   **Ashwani Bhat**
Ayush Jain   **Atin Vikram Singh**   **Ashutosh Modi**
Indian Institute of Technology Kanpur (IIT-K)
`{ajoshi, ashutoshm}@cse.iitk.ac.in`
`{ashubhat44}@gmail.com`
`{aayushj, atinvs}@iitk.ac.in`

Abstract

Emotions are an inherent part of human interactions, and consequently, it is imperative to develop AI systems that understand and recognize human emotions. During a conversation involving various people, a person's emotions are influenced by the other speaker's utterances and their own emotional state over the utterances. In this paper, we propose COntextualized Graph Neural Network based Multimodal Emotion recognitioN (COGMEN) system that leverages local information (i.e., inter/intra dependency between speakers) and global information (context). The proposed model uses Graph Neural Network (GNN) based architecture to model the complex dependencies (local and global information) in a conversation. Our model gives state-of-the-art (SOTA) results on IEMOCAP and MOSEI datasets, and detailed ablation experiments show the importance of modeling information at both levels.

1 Introduction

Emotions are intrinsic to humans and guide their behavior and are indicative of the underlying thought process (Minsky, 2007). Consequently, understanding and recognizing emotions is vital for developing AI technologies (e.g., personal digital assistants) that interact directly with humans. During a conversation between a number of people, there is a constant ebb and flow of emotions experienced and expressed by each person. The task of multimodal emotion recognition addresses the problem of monitoring the emotions expressed (via various modalities, e.g., video (face), audio (speech)) by individuals in different settings such as conversations.

Emotions are physiological, behavioral, and communicative reactions to cognitively processed stimuli (Planalp et al., 2018). Emotions are often a result of internal physiological changes, and
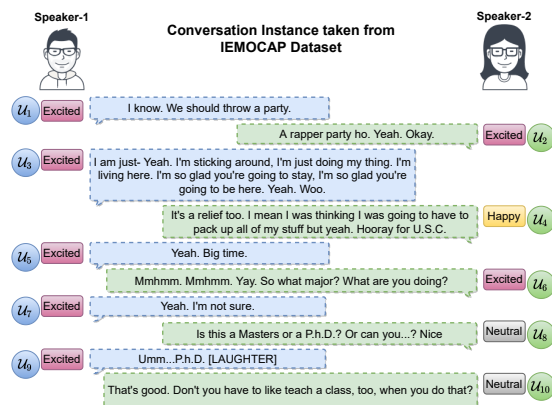

Figure 1: An example conversation between two speakers, with corresponding emotions evoked for each utterance.

these physiological reactions may not be noticeable by others and are therefore intra-personal. For example, in a conversational setting, an emotion may be a communicative reaction that has its origin in a sentence spoken by another person, acting as a stimulus. The emotional states expressed in utterances correlate with the context directly; for example, if the underlying context is about a happy topic like celebrating a festival or description of a vacation, there will be more positive emotions like joy and surprise. Consider the example shown in Figure 1, where the context depicts an exciting conversation. Speaker-1 being excited about his admission affects the flow of emotions in the entire context. The emotion states of Speaker-2 show the dependency on Speaker-1 in $\mathcal{U}_2, \mathcal{U}_4$ and $\mathcal{U}_6$, and maintains intra-personal state depicted in $\mathcal{U}_8$ and $\mathcal{U}_{10}$ by being curious about the responses of Speaker-1. The example conversation portrays the effect of global information as well as inter and intra dependency of speakers on the emotional states of the utterances. Moreover, emotions are a multimodal phenomenon; a person takes cues from different modalities (e.g., audio, video) to infer the emotions of others, since, very often, the in-

4148

formation in different modalities complement each other. In this paper, we leverage these intuitions and propose **COGMEN**: **CO**ntextualized **G**raph neural network based **M**ultimodal **E**motion recognitio**N** architecture that addresses both, the effects of context on the utterances and inter and intra dependency for predicting the per-utterance emotion of each speaker during the conversation. There has been a lot of work on unimodal (using text only) prediction, but our focus is on multimodal emotion prediction. As is done in literature on multimodal emotion prediction, we do not focus on comparison with unimodal models. As shown via experiments and ablation studies, our model leverages both the sources (i.e., local and global) of information to give state-of-the-art (SOTA) results on the multimodal emotion recognition datasets IEMOCAP and MOSEI. In a nutshell, we make the following contributions in this paper:

- We propose a Contextualized Graph Neural Network (GNN) based Multimodal Emotion Recognition architecture for predicting per utterance per speaker emotion in a conversation. Our model leverages both local and global information in a conversation. We use GraphTransformers (Shi et al., 2021) for modeling speaker relations in multimodal emotion recognition systems.

- Our model gives SOTA results on the multimodal Emotion recognition datasets of IEMOCAP and MOSEI.

- We perform a thorough analysis of the model and its different components to show the importance of local and global information along with the importance of the GNN component. We release the code for models and experiments: `https://github.com/Exploration-Lab/COGMEN`

2 Related Work

Emotion recognition is an actively researched problem in NLP (Sharma and Dhall, 2021; Sebe et al., 2005). The broad applications ranging from emotion understanding systems, opinion mining from a corpus to emotion generation have attracted active research interest in recent years (Dhuheir et al., 2021; Franzen et al., 2021; Vinola and Vimaladevi, 2015; Kołakowska et al., 2014; Colombo et al., 2019; Janghorbani et al., 2019; Goswamy et al., 2020; Singh et al., 2021a; Agarwal et al., 2021; Singh et al., 2021b). Availability of bench-mark multimodal datasets, such as CMU-MOSEI (Zadeh et al., 2018b), and IEMOCAP (Busso et al., 2008), have accelerated the progress in the area. Broadly speaking, most of the existing work in this area can be categorized mainly into two areas: *unimodal* approaches and *multimodal* approaches. Unimodal approaches tend to consider the text as a prominent mode of communication and solve the emotion recognition task using only text modality. In contrast, multimodal approaches are more naturalistic and consider multiple modalities (audio+video+text) and fuse them to recognize emotions. In this paper, we propose a multimodal approach to emotion recognition. Nevertheless, we briefly outline some of the prominent unimodal approaches as some of the techniques are applicable to our setting.

Unimodal Approaches: COSMIC (Yu et al., 2019) performs text only emotion classification problem by leveraging commonsense knowledge. DialogXL (Shen et al., 2021a) uses XLnet (Yang et al., 2019) as architecture in dialogue feature extraction. CESTa (Wang et al., 2020) captures the emotional consistency in the utterances using Conditional Random Fields (Lafferty et al., 2001) for boosting the performance of emotion classification. Other popular approaches parallel to our work use graph-based neural networks as their baseline and solve the context propagation issues in RNN-based architectures, including DialogueGCN (Ghosal et al., 2019), RGAT (Ishiwatari et al., 2020), ConGCN (Zhang et al., 2019), and SumAggGin (Sheng et al., 2020). Some of the recent approaches like DAG-ERC (Shen et al., 2021b) combine the strengths of conventional graph-based neural models and recurrence-based neural models.

Multimodal Approaches: Due to the high correlation between emotion and facial cues (Ekman, 1993), fusing modalities to improve emotion recognition has drawn considerable interest (Sebe et al., 2005). Some of the initial approaches include Datcu and Rothkrantz (2014), who fused acoustic information with visual cues for emotion recognition. Wollmer et al. (2010) use contextual information for emotion recognition in a multimodal setting. In the past decade, the growth of deep learning has motivated a wide range of approaches in multimodal settings. The Memory Fusion network (MFN) (Zadeh et al., 2018a) proposes synchronizing multimodal sequences using multi-view gated memory storing intra-view and cross-view interac-

tions through time. Graph-MFN (Bagher Zadeh et al., 2018) extends the idea of MFN and introduces Dynamic Fusion Graph (DFG), which learns to model the n-modal interactions and alter its structure dynamically to choose a fusion graph based on the importance of each n-modal dynamics during inference. Conversational memory network (CMN) (Hazarika et al., 2018b) leverages contextual information from the conversation history and uses gated recurrent units to model past utterances of each speaker into memories. Tensor fusion Network (TFN) (Zadeh et al., 2017) uses an outer product of the modalities. Other popular approaches include DialogueRNN (Majumder et al., 2019) that proposes an attention mechanism over the different utterances and models emotional dynamics by its party GRU and global GRU. B2+B4 (Kumar and Vepa, 2020), use a conditional gating mechanism to learn cross-modal information. bc-LSTM (Poria et al., 2017) proposes an LSTM-based model that captures contextual information from the surrounding utterances. Multilogue-Net (Shenoy and Sardana, 2020) proposes a solution based on a context-aware RNN and uses pairwise attention as a fusion mechanism for all three modalities (audio, video, and text). Recently, Delbrouck et al. (2020) proposed TBJE, a transformer-based architecture with modular co-attention (Yu et al., 2019) to encode multiple modalities jointly. CONSK-GCN (Fu et al., 2021) uses graph convolutional network (GCN) with knowledge graphs. Lian et al. (2020) use GNN based architecture for Emotion Recognition using text and speech modalities. Af-CAN (Wang et al., 2021a) proposes RNN based on contextual attention for modeling the transaction and dependence between speakers.

3 Proposed Model

In a conversation involving different speakers, there is a continuous ebb and flow in the emotions of each of the speakers, usually triggered by the context and reactions of other speakers. Inspired by this intuition, we propose a multimodal emotion prediction model that leverages contextual information, inter-speaker and intra-speaker relations in a conversation.

In our model, we leverage both the context of dialogue and the effect of nearby utterances. We model these two sources of information via two means: **1) Global Information**: How to capture the impact of underlying context on the emotional

state of an utterance? **2) Local information**: How to establish relations between the nearby utterances that preserve both inter-speaker and intra-speaker dependence on utterances in a dialogue?

Global Information: We want to have a unified model that can capture the underlying context and handle its effect on each utterance present in the dialogue. A transformer encoder (Vaswani et al., 2017) architecture is a suitable choice for this goal. Instead of following the conventional sequential encoding by adding positional encodings to the input, in our approach, a simple transformer encoder without any positional encodings leverages the entire context to generate distributed representations (features) efficiently corresponding to each utterance. The transformer facilitates the flow of information from all utterances when predicting emotion for a particular utterance.

Local Information: The emotion expressed in an utterance is often triggered by the information in neighboring utterances. We establish relations between the nearby utterances in a way that is capable of capturing both inter-speaker and intra-speaker effects of stimulus over the emotion state of an utterance. Our approach comes close to DialogueGCN (Ghosal et al., 2019), and we define a graph where each utterance is a node, and directed edges represent various relations. We define relations (directed edges) between nodes $\mathcal{R}_{ij} = u_i \rightarrow u_j$, where the direction of the arrow represents the spoken order of utterances. We categorize the directed relations into two types, for self-dependent relations between the utterances spoken by the same speaker $\mathcal{R}_{intra}$, and interrelations between the utterances spoken by different speakers $\mathcal{R}_{inter}$. We propose to use Relational GCN (Schlichtkrull et al., 2018) followed by a GraphTransformer (Shi et al., 2021) to capture dependency defined by the relations.

3.1 Overall Architecture

Figure 2 shows the detailed architecture. The input utterances go as input to the **Context Extractor** module, which is responsible for capturing the global context. The features extracted for each utterance by the context extractor form a graph based on interactions between the speakers. The graph goes as input to a **Relational GCN**, followed by **GraphTransformer**, which uses the formed graph to capture the inter and intra-relations between the utterances. Finally, two linear layers acting as an **emotion classifier** use the features obtained for all
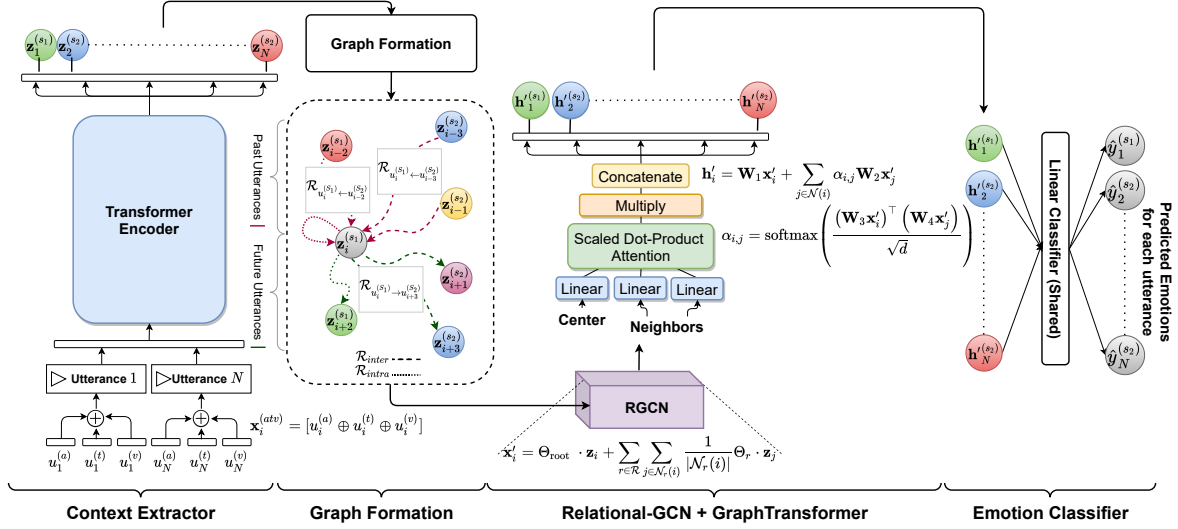
Figure 2: The proposed model (**COGMEN**) architecture.

the utterances to predict the corresponding emotions.

Context Extractor: Context Extractor takes concatenated features of multiple modalities (audio, video, text) as input for each dialogue utterance $(u_i; i = 1, \ldots, n)$ and captures the context using a transformer encoder. The feature vector for an utterance u_i with the input features corresponding to available modalities, audio ($u_i^{(a)} \in \mathbb{R}^{d_a}$), text ($u_i^{(t)} \in \mathbb{R}^{d_t}$) and video ($u_i^{(v)} \in \mathbb{R}^{d_v}$) is:

$$\mathbf{x}_i^{(atv)} = [u_i^{(a)} \oplus u_i^{(t)} \oplus u_i^{(v)}] \in \mathbb{R}^d$$

where $d = d_a + d_t + d_v$. The combined features matrix for all utterances in a dialogue is given by:

$$\mathbf{X} = \mathbf{x}^{(atv)} = [\mathbf{x}_1^{(atv)}, \mathbf{x}_2^{(atv)} \ldots, \mathbf{x}_n^{(atv)}]^T$$

We define a Query, a Key, and a Value vector for encoding the input features $\mathbf{X} \in \mathbb{R}^{n \times d}$ as follows:

$$Q^{(h)} = \mathbf{X} W_{h,q},$$
$$K^{(h)} = \mathbf{X} W_{h,k},$$
$$V^{(h)} = \mathbf{X} W_{h,v},$$

where, $W_{h,q}, W_{h,k}, W_{h,v} \in \mathbb{R}^{d \times k}$

The attention mechanism captures the interaction between the Key and Query vectors to output an attention map $\alpha^{(h)}$, where σ_j denotes the softmax function over the row vectors indexed by j:

$$\alpha^{(h)} = \sigma_j \left(\frac{Q^{(h)} (K^{(h)})^T}{\sqrt{k}} \right)$$

where $\alpha^{(h)} \in \mathbb{R}^{n \times n}$ represents the attention weights for a single attention head (h). The obtained attention map is used to compute a weighted sum of the values for each utterance:

$$\text{head}^{(h)} = \alpha^{(h)} (V^{(h)}) \in \mathbb{R}^{n \times k}$$
$$\mathbf{U}' = [\text{head}^{(1)} \oplus \text{head}^{(2)} \oplus \ldots \text{head}^{(H)}] W^o$$

where, $W^o \in \mathbb{R}^{kH \times d}$ and H represents the total number of heads in multi-head attention. Note $\mathbf{U}' \in \mathbb{R}^{n \times d}$. We add residual connection $\mathbf{X}$ and apply LayerNorm, followed by a feed forward and Add & Norm layer:

$$\mathbf{U} = \text{LayerNorm} \left(\mathbf{X} + \mathbf{U}'; \gamma_1, \beta_1 \right);$$
$$\mathbf{Z}' = \text{ReLU} \left(\mathbf{U} W_1 \right) W_2;$$
$$\mathbf{Z} = \text{LayerNorm} \left(\mathbf{U} + \mathbf{Z}'; \gamma_2, \beta_2 \right);$$

where, $\gamma_1, \beta_1 \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{d \times m}$, $W_2 \in \mathbb{R}^{m \times d}$, and $\gamma_2, \beta_2 \in \mathbb{R}^d$. The transformer encoder provides features corresponding to every utterance in a dialogue ($[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^T = \mathbf{Z} \in \mathbb{R}^{n \times d}$).

Graph Formation: A graph captures inter and intra-speaker dependency between utterances. Every utterance acts as a node of a graph that is connected using directed relations (past and future relations). We define relation types as speaker to speaker. Formally, consider a conversation between M speakers defined as a dialogue $\mathcal{D} = \{\mathcal{U}^{S_1}, \mathcal{U}^{S_2}, \ldots, \mathcal{U}^{S_M}\}$, where $\mathcal{U}^{S_1} = \{u_1^{(S_1)}, u_2^{(S_1)}, \ldots, u_n^{(S_1)}\}$ represent the set of utterances spoken by speaker-1. We define intra relations between the utterances spoken by the same

speaker, $\mathcal{R}_{intra} \in \{\mathcal{U}^{S_i} \to \mathcal{U}^{S_i}\}$, and inter relations between the utterances spoken by different speakers, $\mathcal{R}_{inter} \in \{\mathcal{U}^{S_i} \to \mathcal{U}^{S_j}\}_{i \neq j}$. We further consider a window size and use $\mathcal{P}$ and $\mathcal{F}$ as hyperparameters to form relations between the past $\mathcal{P}$ utterances and future $\mathcal{F}$ utterances for every utterance in a dialogue. For instance, $\mathcal{R}_{intra}$ and $\mathcal{R}_{inter}$ for utterance $u_i^{(S_1)}$ (spoken by speaker-1) are defined as:

$$\mathcal{R}_{intra}(u_i^{(S_1)}) = \{\, u_i^{(S_1)} \leftarrow u_{i-\mathcal{P}}^{(S_1)} \ldots u_i^{(S_1)} \leftarrow u_{i-1}^{(S_1)},$$
$$u_i^{(S_1)} \leftarrow u_i^{(S_1)}, u_i^{(S_1)} \to u_{i+1}^{(S_1)} \ldots u_i^{(S_1)} \to u_{i+\mathcal{F}}^{(S_1)} \}$$
$$\mathcal{R}_{inter}(u_i^{(S_1)}) = \{\, u_i^{(S_1)} \leftarrow u_{i-\mathcal{P}}^{(S_2)}, \ldots, u_i^{(S_1)} \leftarrow u_{i-1}^{(S_2)},$$
$$u_i^{(S_1)} \to u_{i+1}^{(S_2)}, \ldots, u_i^{(S_1)} \to u_{i+\mathcal{F}}^{(S_2)} \}$$

where $\leftarrow$ and $\rightarrow$ represent the past and future relation type respectively (example in Appendix F).

Relational Graph Convolutional Network (RGCN): The vanilla RGCN (Schlichtkrull et al., 2018) helps accumulate relation-specific transformations of neighboring nodes depending on the type and direction of edges present in the graph through a normalized sum. In our case, it captures the inter-speaker and intra-speaker dependency on the connected utterances.

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

where $\mathcal{N}_r(i)$ denotes the set of neighbor indices of node i under relation $r \in \mathcal{R}$, Θ_{root} and Θ_r denote the learnable parameters of RGCN, $|\mathcal{N}_r(i)|$ is the normalization constant and $\mathbf{z}_j$ is the utterance level feature coming from the transformer.

GraphTransformer: For extracting rich representation from the node features, we use a GraphTransformer (Shi et al., 2021). GraphTransformer adopts the vanilla multi-head attention into graph learning by taking into account nodes connected via edges. Given node features $H = \mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ obtained from RGCN,

$$\mathbf{h}_i' = \mathbf{W}_1 \mathbf{x}_i' + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \mathbf{W}_2 \mathbf{x}_j'$$

where the attention coefficients $\alpha_{i,j}$ are computed via multi-head dot product attention:

$$\alpha_{i,j} = \text{softmax}\left(\frac{(\mathbf{W}_3 \mathbf{x}_i')^\top (\mathbf{W}_4 \mathbf{x}_j')}{\sqrt{d}} \right)$$

Dataset	Number of dialogues [utterances]		
	train	valid	test
IEMOCAP	120 [5810 (5146+664)]		31 [1623]
MOSEI	2249 [16327]	300 [1871]	646 [4662]

Table 1: Dataset Statistics.

Emotion Classifier: A linear layer over the features extracted by GraphTransformer ($\mathbf{h}_i'$) predicts the emotion corresponding to the utterance.

$$h_i = \text{ReLU}(W_1 \mathbf{h}_i' + b_1)$$
$$\mathcal{P}_i = \text{softmax}(W_2 h_i + b_2)$$
$$\hat{y}_i = \arg\max(\mathcal{P}_i)$$

where $\hat{y}_i$ is the emotion label predicted for the utterance u_i.

4 Experiments

We experiment for the Emotion Recognition task on the two widely used datasets: **IEMOCAP** (Busso et al., 2008) and **MOSEI** (Zadeh et al., 2018b). The dataset statistics are given in Table 1. IEMOCAP is a dyadic multimodal emotion recognition dataset where each utterance in a dialogue is labeled with one of the six emotion categories: anger, excited, sadness, happiness, frustrated, and neutral. In literature, two IEMOCAP settings are used for testing, one with 4 emotions (anger, sadness, happiness, neutral) and one with 6 emotions. We experiment with both of these settings. MOSEI is a multimodal emotion recognition dataset annotated with 7 sentiments (-3 (highly negative) to +3 (highly positive)) and 6 emotion labels (happiness, sadness, disgust, fear, surprise, and anger). Note that the emotion labels differ across the datasets. We use weighted F1-score and Accuracy as evaluation metrics (details in Appendix C).

Implementation Details: For IEMOCAP, audio features (size 100) are extracted using OpenSmile (Eyben et al., 2010), video features (size 512) are taken from Baltrusaitis et al. (2018), and text features (size 768) are extracted using sBERT (Reimers and Gurevych, 2019). Audio features for the MOSEI dataset were taken from Delbrouck et al. (2020), which are extracted using librosa (McFee et al., 2015) with 80 filter banks, making the feature vector size of 80. The video features (size 35) are taken from Zadeh et al. (2018b). The textual features (size 768) are obtained using sBERT. The textual features are

Models	IEMOCAP: Emotion Categories							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg.	
	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	Acc. (%)	F1 (%)
bc-LSTM	35.6	69.2	53.5	66.3	61.1	62.4	59.8	59.0
memnet	33.0	69.3	55.0	66.1	62.3	63.0	59.9	59.5
TFN	33.7	68.6	55.1	64.2	62.4	61.2	58.8	58.5
MFN	34.1	70.5	52.1	66.8	62.1	62.5	60.1	59.9
CMN	32.6	72.9	56.2	64.6	67.9	63.1	61.9	61.4
ICON	32.8	74.4	60.6	68.2	68.4	**66.2**	64.0	63.5
DialogueRNN	32.8	78.0	59.1	63.3	73.6	59.4	63.3	62.8
CAN	31.8	71.9	60.4	66.7	68.5	66.1	63.2	62.4
Af-CAN	37.0	72.1	60.7	**67.3**	66.5	66.1	64.6	63.7
COGMEN	**51.9**	**81.7**	**68.6**	66.0	**75.3**	58.2	**68.2**	**67.6**

Table 2: Results on IEMOCAP (6-way) multimodal (A+T+V) setting. Avg. denotes weighted average.

sentence-level static features. For Audio and Visual modalities, we use sentence/utterance level features by averaging all the token level features. We fuse the features of all the available modalities (A(audio)+T(text)+V(video): ATV) via concatenation. We also explored other fusion mechanisms (Appendix G.1). However, concatenation gave the best performance. We conduct a hyper-parameter search for our proposed model using Bayesian optimization techniques (details in Appendix A).

Baselines: We do a comprehensive evaluation of **COGMEN** by comparing it with a number of baseline models. For IEMOCAP, we compare our model with the existing multimodal frameworks (Table 2), which includes DialogueRNN (Majumder et al., 2019), bc-LSTM (Poria et al., 2017), CHFusion (Majumder et al., 2018), memnet (Sukhbaatar et al., 2015), TFN (Zadeh et al., 2017), MFN (Zadeh et al., 2018a), CMN (Hazarika et al., 2018b), ICON (Hazarika et al., 2018a), and Af-CAN (Wang et al., 2021b). For MOSEI, **COGMEN** is compared (Table 4) with multimodal models, including Multilogue-Net (Shenoy and Sardana, 2020) and TBJE (Delbrouck et al., 2020) (details and analysis of baselines in §6).

Model	F1-score (%)
bc-LSTM	75.13
CHFusion	76.80
COGMEN	**84.50**

Table 3: Results on IEMOCAP dataset for 4 emotion classes in multimodal setting (weighted F1-score).

5   Results and Analysis

IEMOCAP: Table 2 shows the results for IEMO-CAP (6-way) multimodal setting. Overall, **COG-**

MEN performs better than all the previous baselines as measured using accuracy and F1-score. We also see an improvement in the class-wise F1 for happy, sad, neutral, and excited emotions. This improvement is possibly due to the GNN architecture (described in analysis later) that we are using in our model, and none of the previous multimodal baselines uses GNN in their architecture. Results for IEMOCAP (4-way) setting are in Table 3. In this setting, **COGMEN** achieves 7.7% improvement over the previous SOTA model.

MOSEI: For emotion classification across 6 emotion classes, we used two settings (as done in previous works): *Binary Classification* across each emotion label where a separate model is trained for every emotion class, and *Multi-label Classification* in which the sentence is tagged with more than 1 emotion and single model predicts multiple classes. The reason for doing this was that Multilogue-Net provides results on binary classification setting and TBJE provides results on Multi-label setting. We ran both models on these settings. For a fair comparison, we use the same utterance level textual features similar to our setting (extracted from sBERT) and train Multilogue-Net architecture on both the settings. Originally, Multilogue-Net used GloVe embeddings (Pennington et al., 2014) for textual features, and actual results in the paper are different than reported here. For TBJE, we use the features provided by the paper as it uses token-level features. **COGMEN** outperforms (Table 4) the baseline models in most of the cases. For 2 class sentiment classification, **COGMEN** outperforms the previous baselines with the highest accuracy score of 85% for *A+T*. For 7 class, our model shows comparable performance. All the multimodal approaches tend to perform poorly when

Model		Sentiment Class Accuracy(%)		Emotion Class (weighted) F1-score (%)						Multi-label Emotion Class (weighted) F1-score (%)					
		2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilogue-Net	T + A + V	82.88	**44.83**	67.84	65.34	67.03	87.79	74.91	**86.05**	70.6	70.7	74.4	86.0	83.4	87.8
TBJE	T	81.9	44.2	-	-	-	-	-	-	63.4	65.8	75.3	84.0	84.5	81.4
	A + T	82.4	43.91	65.91	70.78	70.86	87.79	82.57	86.04	65.5	67.9	76.0	**87.2**	84.5	86.1
	T + A + V	81.5	44.4	-	-	-	-	-	-	64.0	67.9	74.7	84.0	83.6	86.1
COGMEN	T	84.42	43.50	69.28	70.49	73.04	87.80	83.69	85.83	69.92	72.16	77.34	86.39	**86.00**	88.27
	A + T	**85.00**	44.31	68.39	**73.28**	74.98	88.08	**83.90**	85.35	69.62	72.67	76.93	86.39	85.35	88.21
	T + A + V	84.34	43.90	**70.42**	72.31	**76.20**	**88.17**	83.69	85.28	**72.74**	**73.90**	**78.04**	86.71	85.48	**88.37**

Table 4: Results on MOSEI dataset. For emotion classification, a weighted F1-score is used. For Sentiment Classification, the results are reported using accuracy. 2 class sentiment consists of only positive and negative sentiment. 7 class sentiment consists of sentiments from highly negative (-3) to highly positive (+3). For the cells showing '-', the results were not provided in the paper, and we were not able to reproduce the results since *TBJE* used token level features, and we are using sentence-level features.

# Utterances in Context	F1-score (%)
All Utterances in a dialogue	84.50
10 Utterances in a dialogue	77.43 (↓7.07)
3 Utterances in a dialogue	75.39 (↓9.11)

Table 5: Importance of Context in a dialogue. Experiment performed on IEMOCAP (4-way).

	Modalities	T	A+T	A+T+V
(6 way)	Actual	66.00	65.42	67.63
	w/o GNN	64.34 (↓1.66)	61.69 (↓3.73)	62.96 (↓4.14)
	w/o Relations	60.49 (↓5.51)	65.32 (↓0.10)	62.13 (↓5.50)
(4 way)	Actual	81.55	81.59	84.50
	w/o GNN	81.18 (↓0.37)	80.16 (↓1.43)	80.28 (↓4.22)
	w/o Relations	76.76 (↓4.79)	80.27 (↓1.32)	79.61 (↓4.88)

Table 6: Ablation study on IEMOCAP dataset. All values are F1-score (%). The results shows the importance of GCN layer.

Model	Modality	F1-score (%)
4-way		
DialogueGCN	T	71.58
DialogXL	T	73.02
DAG-ERC	T	78.08
COGMEN	T	**81.55**
	A+T+V	**84.50**
6-way		
EmoBERTa	T	**68.57**
DAG-ERC	T	68.03
CESTa	T	67.10
SumAggGIN	T	66.61
DialogueCRN	T	66.20
DialogXL	T	65.94
DialogueGCN	T	64.18
COGMEN	T	66.00
	A+T+V	67.63

Table 7: Comparison with unimodal architectures on IEMOCAP dataset.

adding visual modality, possibly because of noise present in the visual modality and lack of alignment with respect to other modalities. In contrast, our model can capture rich relations across the modalities and show a performance boost while adding visual modality.

We conducted further analysis on our model. Although due to space limitations, the results below mainly describe experiments over IEMOCAP, similar trends were observed for MOSEI as well.

Effect of Local and Global Info.: We test our architecture in two information utilization settings: global and local. To test the importance of context in our architecture, we create a sub-dataset using the IEMOCAP (4-way) setting by splitting each dialogue into n utterances and training our architecture. Table 5 shows the decrease in performance with number of utterances present in a dialogue (more details on effect of window size in Appendix G.2). This experiment helps understand the importance of context in a dialogue. Moreover, it points towards challenges in developing a real-time system (details in §6). We test the local information hypothesis by removing the GNN module and directly passing the context extracted features to the emotion classifier. Table 6 shows the drop in performance across modalities when the GNN component is removed from the architecture, making our local information hypothesis more concrete.

Effect of Relation Types: We also test the effect of inter and intra-relations in the dialogue graph by making all relations of the same type and training the architecture. We observe a drop in performance (Table 6) when the relations are kept the same in the

graph formation step. The explicit relation formation helps capture the local dependencies present in the dialogue.

Effect of Modalities: The focus of this work is multimodal emotion recognition. However, just for the purpose of comparison, we also compare with unimodal (text only) approaches. We compare (Table 7) with EmoBERTa (Kim and Vossen, 2021), DAG-ERC (Shen et al., 2021b), CESTa (Wang et al., 2020), SumAggGIN (Sheng et al., 2020), DialogueCRN (Hu et al., 2021), DialogXL (Shen et al., 2021a) and DialogueGCN (Ghosal et al., 2019). Text-based models are specifically optimized for text modalities and incorporate changes to architectures to cater to text. It is not fair to compare with our multimodal approach from that perspective. As shown in results, **COGMEN**, being a fairly generic architecture, still gives better (for IEMOCAP (4-way)) or comparable performance with respect to the SOTA unimodal architectures. In the case of our model, adding more information via other modalities helps to improve the performance. Results on different modality combinations are in Appendix D.
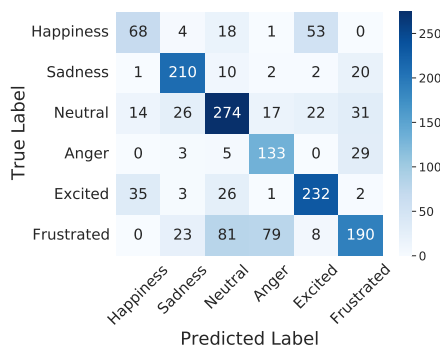

Figure 3: Confusion Matrix for IEMOCAP (6-way)

Error Analysis: After analysing the predictions made across the datasets, we find that our model falls short in distinguishing between similar emotions, such as *happiness* vs *excited* and *anger* vs *frustration* (Figure 3). This issue also exists in previous methods as reported in Shen et al. (2021b), and Ghosal et al. (2019). We also find that our model misclassifies the other emotion labels as neutral because of a more significant proportion of neutral labeled examples. Moreover, we observe the accuracy of our model when classifying examples having emotion shift is 53.6% compared to 74.2% when the emotion remains the same (more details in Appendix B).
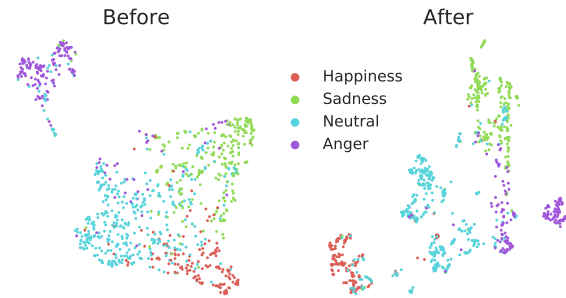
Efficacy of the GNN Layer: For observing the

Figure 4: UMAP (Becht et al., 2019) representation of IEMOCAP (4-way) features before and after GNN.

effect of the GNN component in our architecture, we also visualize the features before and after the GNN component. Figure 4 clearly shows the better formation of emotion clusters depicting the importance of capturing local dependency in utterances for better performance in emotion recognition (more in Appendix E and Appendix Figure-9).

Importance of utterances: To verify the effect of utterances and their importance in a prediction for a dialogue, we infer the trained model on dialogues by masking one utterance at a time and calculating the F1-score for prediction. Figure 5 shows the obtained results for a dialogue (Appendix Table 10) instance taken randomly from IEMOCAP (4-way) (more in Appendix E). For the first 4 utterances, emotions state being neutral, the effect of masking the utterances is significantly less. In contrast, masking the utterances with emotion shift (9, 10, 11) completely drops the dialogue's F1-score, showing that our architecture captures the effects of emotions present in the utterances.
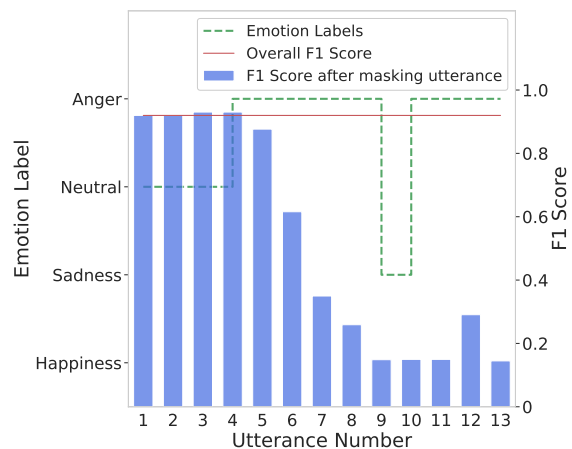

Figure 5: Importance of utterances in IEMOCAP (4-way). Performance drop is observed while masking 9th, 10th and 11th utterances during inference.

6 Discussion

Comparison with Baselines: Emotion recognition in a multimodal conversation setting comes with two broadly portrayed research challenges (Poria et al., 2019), first, the ability of a model to capture global and local context present in the dialogues, and second, the ability to maintain self and interpersonal dependencies among the speakers. All the popular baselines like Dialogue-GCN (Ghosal et al., 2019), DialogueRNN (Majumder et al., 2019), bc-LSTM (Poria et al., 2017) Af-CAN (Wang et al., 2021a), etc., try to address these challenges by proposing various architectures. bc-LSTM (bi-directional contextual LSTM (Poria et al., 2017)) uses LSTM to capture the contextual information and maintain long relations between the utterances from the past and future. Another contemporary architecture Af-CAN (Wang et al., 2021a) utilizes recurrent neural networks based on contextual attention to model the interaction and dependence between speakers and uses bi-directional GRU units to capture the global features from past and future. We propose to address these issues using a unified architecture that captures the effect of context on utterances while maintaining the states for self and interpersonal dependencies. We make use of transformers for encoding the global context and make use of GraphTransformers to capture the self and interpersonal dependencies. Our way of forming relational graphs between the utterances comes close to DialogueGCN (unimodal architecture). We further use a shared Emotion classifier for predicting emotions from all the obtained utterance level features. Moreover, our unified architecture handles multiple modalities effectively and shows an increase in performance after adding information from other modalities.

Limitations (Offline Setting): A noteworthy limitation of all the proposed Emotion Recognition approaches (including the current one) is that they use global context from past and future utterances to predict emotions. However, baseline systems compared in this paper are also offline systems. For example, bc-LSTM (bi-directional contextual LSTM) and Af-CAN use utterances from the past and future to predict emotions. Other popular baselines like DialogueGCN and DialogueRNN (BiDialogueRNN) also peek into the future, assuming the presence of all the utterances during inference (offline setting). All such systems that depend on future information can only be used in an offline

setting to process and tag the dialogue. An Emotion Recognition system that could work in an online setting exhibits another line of future work worth exploring due to its vast use cases in live telecasting and telecommunication. A possible approach to maintain the context in an online setting would be to take a buffer of smaller context size, where the model can predict emotions taking not the complete dialogue but a smaller subset of it as input in real-time. We tried exploring this setting for our architecture with an online buffer of maintaining a smaller context window. For experimenting with it, we created a sub-dataset using the IEMOCAP (4-way) setting by splitting each dialogue into n utterances and training our architecture. Our results in Table 5 show the decrease in performance with the number of utterances present in a dialogue depicting the importance of context in a conversation. Performance improvements in these settings where the system can work in real-time are worth exploring and are an interesting direction for future research.

7 Conclusion and Future Work

We present a novel approach of using GNNs for multimodal emotion recognition and propose **COGMEN: COntextualized GNN based Multimodal Emotion recognitioN**. We test **COGMEN** on two widely known multimodal emotion recognition datasets, IEMOCAP and MOSEI. **COGMEN** outperforms the existing state-of-the-art methods in multimodal emotion recognition by a significant margin (i.e., 7.7% F1-score increase for IEMOCAP (4-way)). By comprehensive analysis and ablation studies over **COGMEN**, we show the importance of different modules. **COGMEN** fuses information effectively from multiple modalities to improve the performance of emotion prediction tasks. We perform a detailed error analysis and observe that the misclassifications are mainly between the similar classes and emotion shift cases. We plan to address this in future work, where the focus will be to incorporate a component for capturing the emotional shifts for fine-grained emotion prediction.

8 Acknowledgements

References

Harsh Agarwal, Keshav Bansal, Abhinav Joshi, and Ashutosh Modi. 2021. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. *CoRR*, abs/2112.01938.

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, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

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(4):335–359.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Comet.ML. 2021. Comet.ML home page.

Dragos Datcu and Leon JM Rothkrantz. 2014. Semantic audio-visual data fusion for automatic emotion recognition. *Emotion recognition: a pattern analysis approach*, pages 411–435.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.

Marwan Dhuheir, Abdullatif Albaseer, Emna Baccour, Aiman Erbad, Mohamed Abdallah, and Mounir Hamdi. 2021. Emotion recognition for healthcare surveillance systems using neural networks: A survey.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *CoRR*, abs/1903.02428.

Marc Franzen, Michael Stephan Gresser, Tobias Müller, and Prof. Dr. Sebastian Mauser. 2021. Developing emotion recognition for video conference software to support people with autism.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. Consk-gcn: Conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

M. Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings*

of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *ACL/IJCNLP (1)*, pages 7042–7052. Association for Computational Linguistics.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Sepehr Janghorbani, Ashutosh Modi, Jakob Buhmann, and Mubbasir Kapadia. 2019. Domain authoring assistant for intelligent virtual agent. AAMAS '19, page 104–112, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv e-prints*, page arXiv:2108.12009.

Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michal R Wrobel. 2014. Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 51–62. Springer.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms

and graph neural networks. In *INTERSPEECH*, pages 2347–2351.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133.

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

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python.

Marvin Minsky. 2007. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. SIMON & SCHUSTER.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sally Planalp, Julie Fitness, and Beverley A. Fehr. 2018. *The Roles of Emotion in Relationships*, 2 edition, Cambridge Handbooks in Psychology, page 256–268. Cambridge University Press.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*, pages 593–607, Cham. Springer International Publishing.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2005. Multimodal approaches for emotion recognition: A survey. *Proceedings of SPIE - The International Society for Optical Engineering*, 5670:56–67. Proceedings of SPIE-IS and T Electronic Imaging - Internet Imaging VI ; Conference date: 18-01-2005 Through 20-01-2005.

Garima Sharma and Abhinav Dhall. 2021. *A Survey on Automatic Multimodal Emotion Recognition in the Wild*, pages 35–64.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. 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, Online. Association for Computational Linguistics.

Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. 2020. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4153–4163.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA. Association for Computational Linguistics.

Yunsheng Shi, Zhengjie Huang, Wenjin Wang, Hui Zhong, Shikun Feng, and Yu Sun. 2021. Masked label prediction: Unified massage passing model for semi-supervised classification. In *IJCAI*.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021a. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91, Online. Association for Computational Linguistics.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2021b. Fine-Grained Emotion Prediction by Modeling Emotion Definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2440–2448, Cambridge, MA, USA. MIT Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

C Vinola and K Vimaladevi. 2015. A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 14(2):24–44.

Tana Wang, Yaqing Hou, Dongsheng Zhou, and Qiang Zhang. 2021a. A contextual attention network for multimodal emotion recognition in conversation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Tana Wang, Yaqing Hou, Dongsheng Zhou, and Qiang Zhang. 2021b. A contextual attention network for multimodal emotion recognition in conversation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195.

Martin Wollmer, Angeliki Metallinou, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multiview sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. 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.

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 *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421. ijcai.org.

Appendix

A Hyperparameter Setting

Hyperparameters used to train our model are described in Table 8 for IEMOCAP (4-way and 6-way) and Table 9 for MOSEI dataset.

Dropout	GNNHead	SeqContext	ILR
0.1	7	4	1e-4

Table 8: **Hyperparameter values for our model on IEMOCAP dataset.** *ILR: Initial learning rate.*

Modalities	Dropout	GNNHead	SeqContext	ILR
T	0.399	3	5	3.3e-3
A+T	0.103	1	2	6.9e-3
A+T+V	0.337	2	1	1.1e-3

Table 9: **Hyperparameter value on MOSEI dataset**. *ILR: Initial learning rate.*

We use PyTorch (Paszke et al., 2019) for training our architecture and PyG (PyTorch Geometric) (Fey and Lenssen, 2019) for the GNN component in our architecture. We use comet (Comet.ML, 2021) for logging all our experiments and its Bayesian optimizer for hyperparameter tuning. Our architecture trained on the IEMOCAP dataset has 55,932,052 parameters and takes around 7 minutes to train for 50 epochs on the NVIDIA Tesla K80 GPU. Comparison of the model with baselines in terms of the number of parameters is challenging, as the baselines parameters vary depending on the hyperparameter setting. Moreover, many baselines do not provide information about the number of parameters.

B Dataset Analysis

We study IEMOCAP dataset in detail for error analysis of our model. We observe the emotion transition at Utterance level (Figure 6) and Speaker level (Figure 7). We find a high percentage of transitions between similar emotions, causing the models to confuse between the similar classes of emotion. Considering the emotion transition between states that are opposite, like from happy to sad, we deduce the poor performance of emotion recognition architectures for such cases. We plan to address this issue in future work where we target a model which performs better in fine-grained emotion recognition and is robust towards the shifts in emotions.
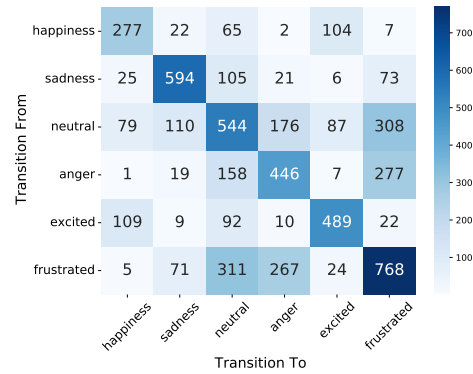

Figure 6: Utterance-level Emotion transition for IEMOCAP. These are emotions transitions in consecutive utterances across speakers.
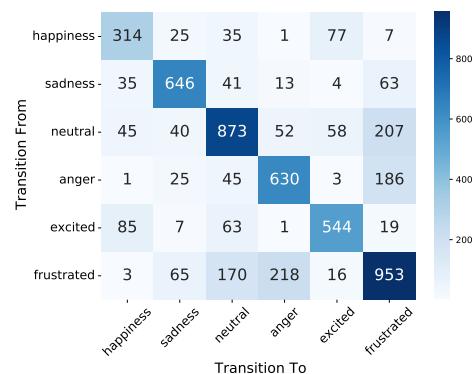

Figure 7: Speaker-level Emotion transition for IEMOCAP. These are emotions transitions in the consecutive utterances of the same speaker.

C Evaluation Metrics

Weighted F1 Score: The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

For weighted F1 score, we calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). **Accuracy**: It is defined as the percentage of correct predictions in the test set.

D Results on Modality Combinations

Table 11 shows results on the IEMOCAP dataset for all the modality combinations for our architectures. Figure 8 shows the confusion matrix for prediction on IEMOCAP 4-way dataset.

4161

Speaker	Utterance Text	Emotion
M	'Why does that bother you?'	neutral
F	"She's been in New York three and a half years. Why all of the sudden?"	neutral
M	'Well maybe. Maybe she just wanted to see her again.'	neutral
F	"What did you mean? He lived next door to the girl all of his life, why wouldn't he want to see her again? Don't look at me like that, he didn't tell me any more than he told you."	neutral
M	"She's not his girl. She knows she's not."	angry
F	"I want you to pretend like he's coming back!"	angry
M	"Because if he's not coming back, then I'll kill myself."	angry
F	'Laugh. Laugh at me, but what happens the night that she goes to sleep in his bed, and his memorial breaks in pieces?"	angry
M	'Only last week, another boy turned up in Detroit, been missing longer than Larry, you read it yourself, '	angry
F	"You've got to believe. You've got to–"	sad
M	"What do you mean me above all? Look at you. You're shaking!"	angry
F	"I can't help it!"	angry
M	'What have I got to hide? What the hell is the matter with you, Kate?'	angry

Table 10: Dialogue utterances corresponding to plot shown in Figure 5.

Modalities	IEMOCAP-4way F1 Score (%)	IEMOCAP-6way F1 Score (%)
a	63.58	47.57
t	81.55	66.00
v	43.85	37.58
at	81.59	65.42
av	64.48	52.20
tv	81.52	62.19
atv	**84.50**	**67.63**

Table 11: Results on IEMOCAP-4way and IEMOCAP-6way datasets

Figure 9: UMAP (Becht et al., 2019) representation of IEMOCAP 6-way features before and after GNN.

Figure 8: Confusion Matrix for IEMOCAP 4-Way classification

Figure 10: Importance of utterances in IEMOCAP classification.
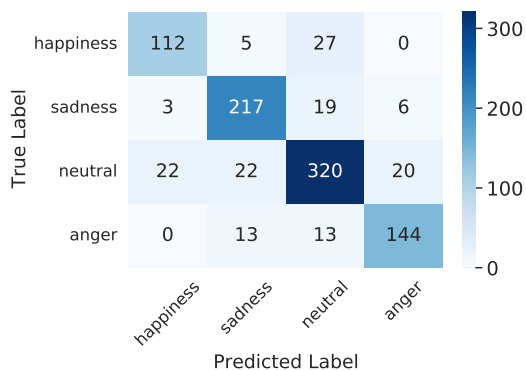
E Additional Analysis

Efficacy of the GNN Layer: We observe the efficacy of the GNN component in our architecture and visualize the features before GNN and after the GNN component (Figure 9) explained in section 5. **Importance of utterances:** Figure 10 shows the obtained results for a dialogue instance taken randomly from IEMOCAP 4-way. For the first 15
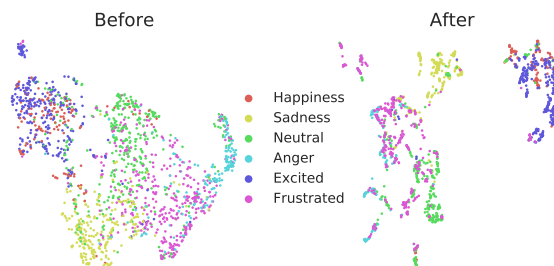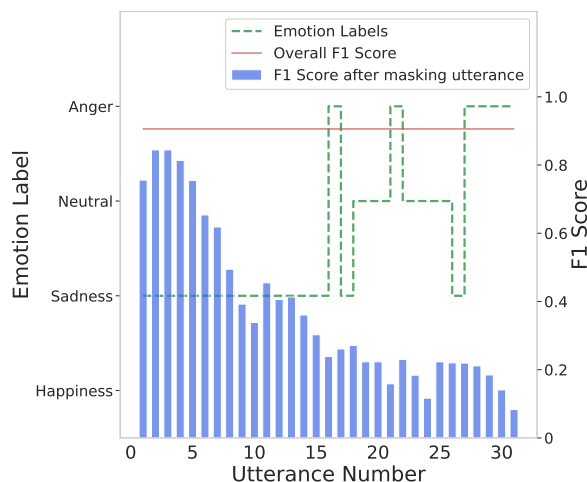
utterances, emotions state being sadness, the effect of masking the utterances is more negligible for the first 5 utterances. This drop depicts the importance of utterances 5-15 that affect future utterances. Further, masking the utterances with high
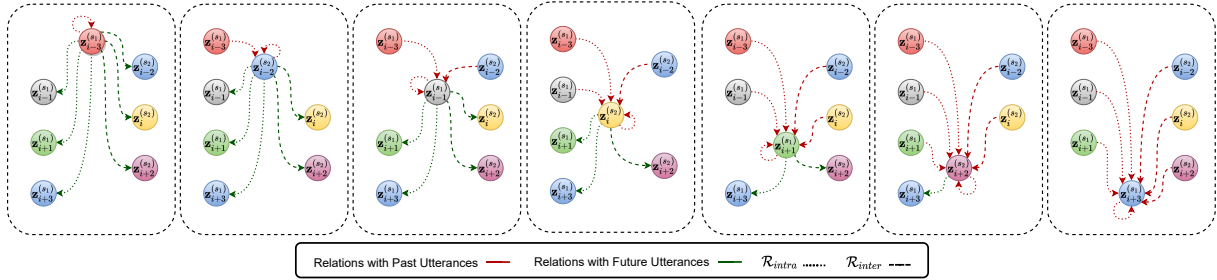
Figure 11: Graph formation process in (**COGMEN**) architecture.

Figure	Central Node	$\mathcal{R}_{intra}$	$\mathcal{R}_{inter}$
(a)	$u_{i-3}^{(S_1)}$	$u_{i-3}^{(S_1)} \leftarrow u_{i-3}^{(S_1)}, u_{i-3}^{(S_1)} \rightarrow u_{i-1}^{(S_1)}, u_{i-3}^{(S_1)} \rightarrow u_{i+1}^{(S_1)}, u_{i-3}^{(S_1)} \rightarrow u_{i+3}^{(S_1)}$	$u_{i-3}^{(S_1)} \rightarrow u_{i-2}^{(S_2)}, u_{i-3}^{(S_1)} \rightarrow u_{i}^{(S_2)}, u_{i-3}^{(S_1)} \rightarrow u_{i+2}^{(S_2)}$
(b)	$u_{i-2}^{(S_2)}$	$u_{i-2}^{(S_2)} \leftarrow u_{i-2}^{(S_2)}, u_{i-2}^{(S_2)} \rightarrow u_{i}^{(S_2)}, u_{i-2}^{(S_2)} \rightarrow u_{i+2}^{(S_2)}$	$u_{i-2}^{(S_2)} \leftarrow u_{i-3}^{(S_1)}, u_{i-2}^{(S_2)} \rightarrow u_{i-1}^{(S_1)}, u_{i-2}^{(S_2)} \rightarrow u_{i+1}^{(S_1)}, u_{i-2}^{(S_2)} \rightarrow u_{i+3}^{(S_1)}$
(c)	$u_{i-1}^{(S_1)}$	$u_{i-1}^{(S_1)} \leftarrow u_{i-3}^{(S_1)}, u_{i-1}^{(S_1)} \leftarrow u_{i-1}^{(S_1)}, u_{i-1}^{(S_1)} \rightarrow u_{i+1}^{(S_1)}, u_{i-1}^{(S_1)} \rightarrow u_{i+3}^{(S_1)}$	$u_{i-1}^{(S_1)} \leftarrow u_{i-2}^{(S_2)}, u_{i-1}^{(S_1)} \rightarrow u_{i}^{(S_2)}, u_{i-1}^{(S_1)} \rightarrow u_{i+2}^{(S_2)}$
(d)	$u_{i}^{(S_2)}$	$u_{i}^{(S_2)} \leftarrow u_{i-2}^{(S_2)}, u_{i}^{(S_2)} \leftarrow u_{i}^{(S_2)}, u_{i}^{(S_2)} \rightarrow u_{i+2}^{(S_2)}$	$u_{i}^{(S_2)} \leftarrow u_{i-3}^{(S_1)}, u_{i}^{(S_2)} \leftarrow u_{i-1}^{(S_1)}, u_{i}^{(S_2)} \rightarrow u_{i+1}^{(S_1)}, u_{i}^{(S_2)} \rightarrow u_{i+3}^{(S_1)}$
(e)	$u_{i+1}^{(S_1)}$	$u_{i+1}^{(S_1)} \leftarrow u_{i-3}^{(S_1)}, u_{i+1}^{(S_1)} \leftarrow u_{i-1}^{(S_1)}, u_{i+1}^{(S_1)} \leftarrow u_{i+1}^{(S_1)}, u_{i+1}^{(S_1)} \rightarrow u_{i+3}^{(S_1)}$	$u_{i+1}^{(S_1)} \leftarrow u_{i-2}^{(S_2)}, u_{i+1}^{(S_1)} \leftarrow u_{i}^{(S_2)}, u_{i+1}^{(S_1)} \rightarrow u_{i+2}^{(S_2)}$
(f)	$u_{i+2}^{(S_2)}$	$u_{i+2}^{(S_2)} \leftarrow u_{i-2}^{(S_2)}, u_{i+2}^{(S_2)} \leftarrow u_{i}^{(S_2)}, u_{i+2}^{(S_2)} \leftarrow u_{i+2}^{(S_2)}$	$u_{i+2}^{(S_2)} \leftarrow u_{i-3}^{(S_1)}, u_{i+2}^{(S_2)} \leftarrow u_{i-1}^{(S_1)}, u_{i+2}^{(S_2)} \leftarrow u_{i+1}^{(S_1)}, u_{i+2}^{(S_2)} \rightarrow u_{i+3}^{(S_1)}$
(g)	$u_{i+3}^{(S_1)}$	$u_{i+3}^{(S_1)} \leftarrow u_{i-3}^{(S_1)}, u_{i+3}^{(S_1)} \leftarrow u_{i-1}^{(S_1)}, u_{i+3}^{(S_1)} \leftarrow u_{i+1}^{(S_1)}, u_{i+3}^{(S_1)} \leftarrow u_{i+3}^{(S_1)}$	$u_{i+3}^{(S_1)} \leftarrow u_{i-2}^{(S_2)}, u_{i+3}^{(S_1)} \leftarrow u_{i}^{(S_2)}, u_{i+3}^{(S_1)} \leftarrow u_{i+2}^{(S_2)}$

Table 12: Relations for each instance of Figure 11, where relations with past utterances are denoted by ($\leftarrow$) and relations with future utterances are denoted by ($\rightarrow$)

Relation Type	Node A	Node B	Relation Causality	Relation
1	$u^{(S_1)}$	$u^{(S_1)}$	Past	$u^{(S_1)} \leftarrow u^{(S_1)}$
2	$u^{(S_1)}$	$u^{(S_2)}$	Past	$u^{(S_1)} \leftarrow u^{(S_2)}$
3	$u^{(S_2)}$	$u^{(S_1)}$	Past	$u^{(S_2)} \leftarrow u^{(S_1)}$
4	$u^{(S_2)}$	$u^{(S_2)}$	Past	$u^{(S_2)} \leftarrow u^{(S_2)}$
5	$u^{(S_1)}$	$u^{(S_1)}$	Future	$u^{(S_1)} \rightarrow u^{(S_1)}$
6	$u^{(S_1)}$	$u^{(S_2)}$	Future	$u^{(S_1)} \rightarrow u^{(S_2)}$
7	$u^{(S_2)}$	$u^{(S_1)}$	Future	$u^{(S_2)} \rightarrow u^{(S_1)}$
8	$u^{(S_2)}$	$u^{(S_2)}$	Future	$u^{(S_2)} \rightarrow u^{(S_2)}$

Table 13: Unique Relation types for a conversation between two speakers

unique relations would be:

$$\begin{aligned}
\text{number of relations} &= 2 \times (S_N)^2 \\
&= 2 \times (2)^2 \\
&= 8
\end{aligned}$$

Table 13 shows the number of possible unique relations for a conversation between two speakers.

G Discussion

G.1 Modality Fusing Mechanisms

While experimenting with the model architecture, we explored various mechanisms for mixing information from multiple modalities. Some of the mechanisms include pairwise attention inspired from Ghosal et al. (2018), bimodal attention present in Multilogue-Net (Shenoy and Sardana, 2020), and crossAttention layer proposed in HKT (Hasan et al., 2021). However, in our case, none of these fusing mechanisms shows significant performance improvement over simple concatenation. Moreover, all these fusing mechanisms require extra computation steps for fusing information. In contrast, a simple concatenation of modality features works well with no additional computational overhead.

emotion shift (15 to 30) drops the F1 score of the dialogue, showing the importance of fluctuations for predicting the emotion states for other utterances.

F Graph Formation

To give a clear picture of the graph formation procedure, we describe the process for utterances spoken in a dialogue. As an illustration, let's consider two speakers, S_1 and S_2, present in a conversation of 7 utterances. Features corresponding to each utterance is shown as a node in Figure 11. Speaker 1 speaks utterances $u_{i-3}, u_{i-1}, u_{i+1}, u_{i+3}$ and Speaker 2 speaks u_{i-2}, u_i, u_{i+2}. After creating the graphs with relations, the constructed graph would look like shown in Figure 11, and the corresponding relations for each instance would be as shown in Table 12. Since there are two speakers in the conversation ($S_N = 2$), the total number of

G.2 Effect of window size in Graph Formation

To explore the effect of window size in the Graph Formation module of our architecture, we conduct experiments with multiple window sizes. The obtained results are present in Table 14. The window size can be treated as a hyperparameter that could be adjusted while training our architecture. Moreover, the freedom of setting the window size makes our architecture more flexible in terms of usage. A larger window size would result in better performance for cases where the inter and intra speaker dependencies are maintained for longer sequences. In contrast, setting a lower window size would be better in a use case where the topic frequently changes in dialogues and speakers are less affected by another speaker. In the future, we plan to explore a dynamic and automatic selection of window size depending on the dialogue instance.

Modalities	Window Past	Window future	F1 Score (%)
atv	1	1	81.72
atv	2	2	83.21
atv	4	4	**84.08**
atv	5	5	83.19
atv	6	6	82.49
atv	7	7	82.28
atv	9	9	82.77
atv	10	10	**84.50**
atv	11	11	83.93
atv	15	15	83.78

Table 14: Results for various window sizes for graph formation on the IEMOCAP (4-way) dataset.