DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation

EMNLP19

Deepanway Ghosal, Navonil Majumder, Soujanya Poria , Niyati Chhaya and Alexander Gelbukh

Singapore University of Technology and Design, Singapore

Instituto Polit´ecnico Nacional, CIC, Mexico

Adobe Research, India

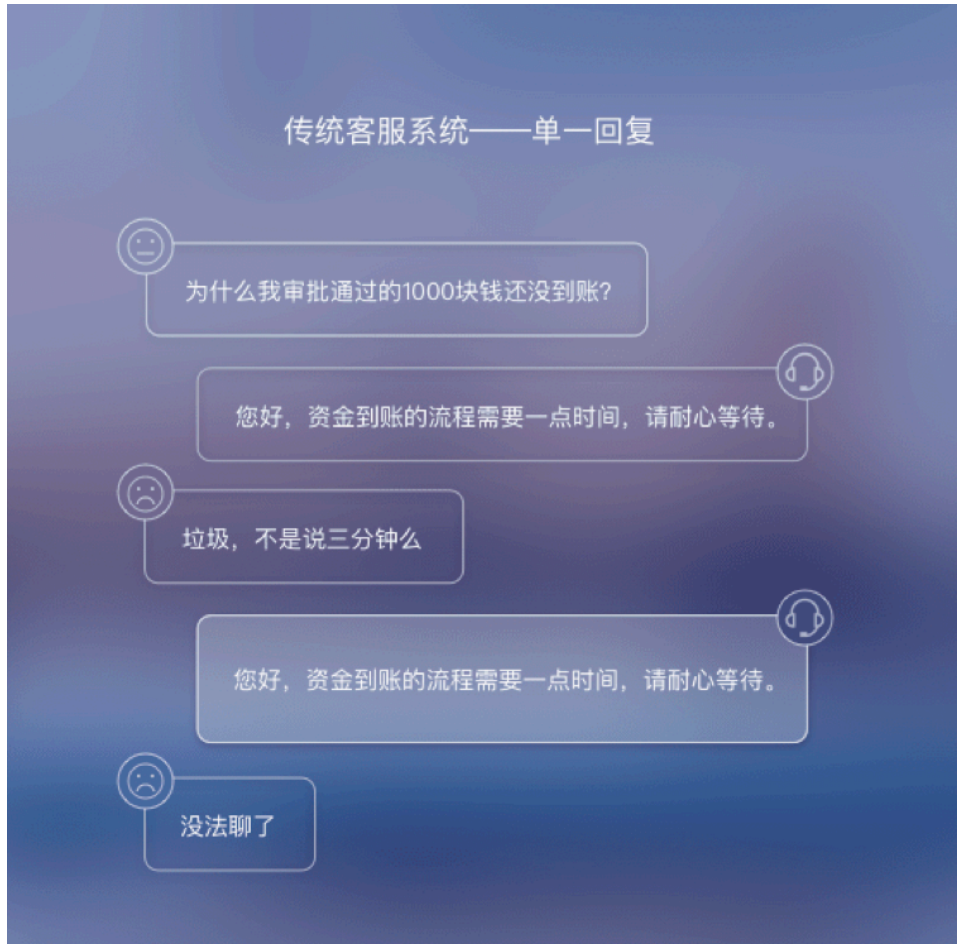
Reporter:XiachongFeng

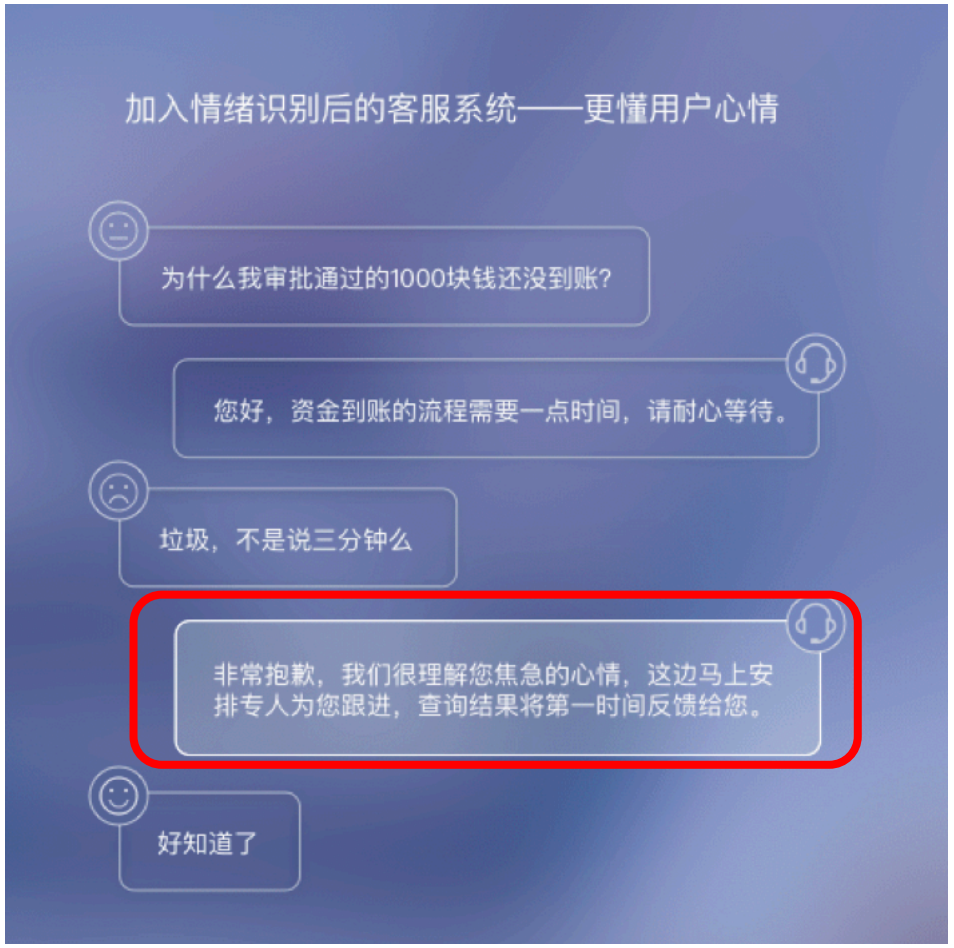
Authors

Deepanway Ghosal

Research fellow in the School of Computer Science & Engineering at NTU Singapore

Emotion recognition in conversation (ERC)





https://ai.baidu.com/tech/nlp/emotion_detection

Emotion recognition in conversation (ERC)

https://www.leiphone.com/news/201805/gRJ1UqPmoCpfHPVL.html

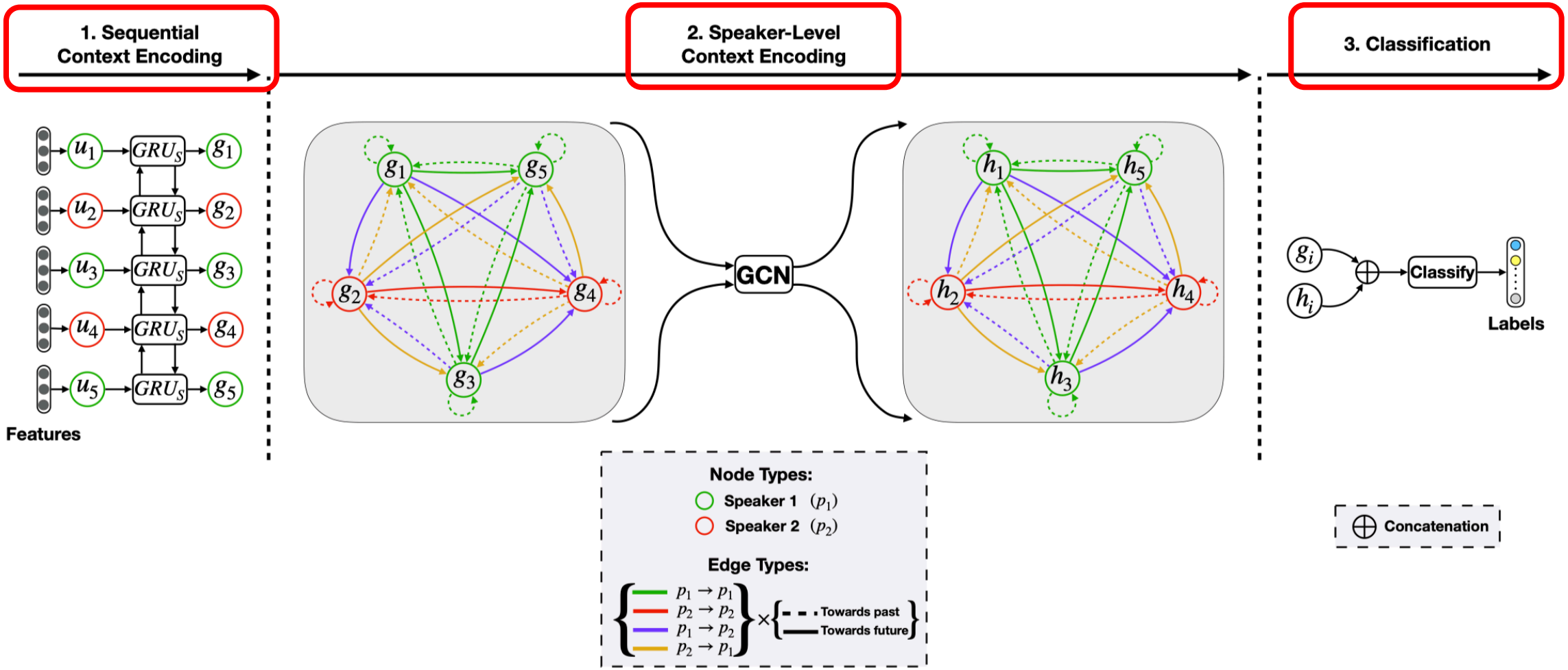
Core Idea

- Leverage **self and inter-speaker dependency** of the interlocutors to model conversational **context** for emotion recognition.

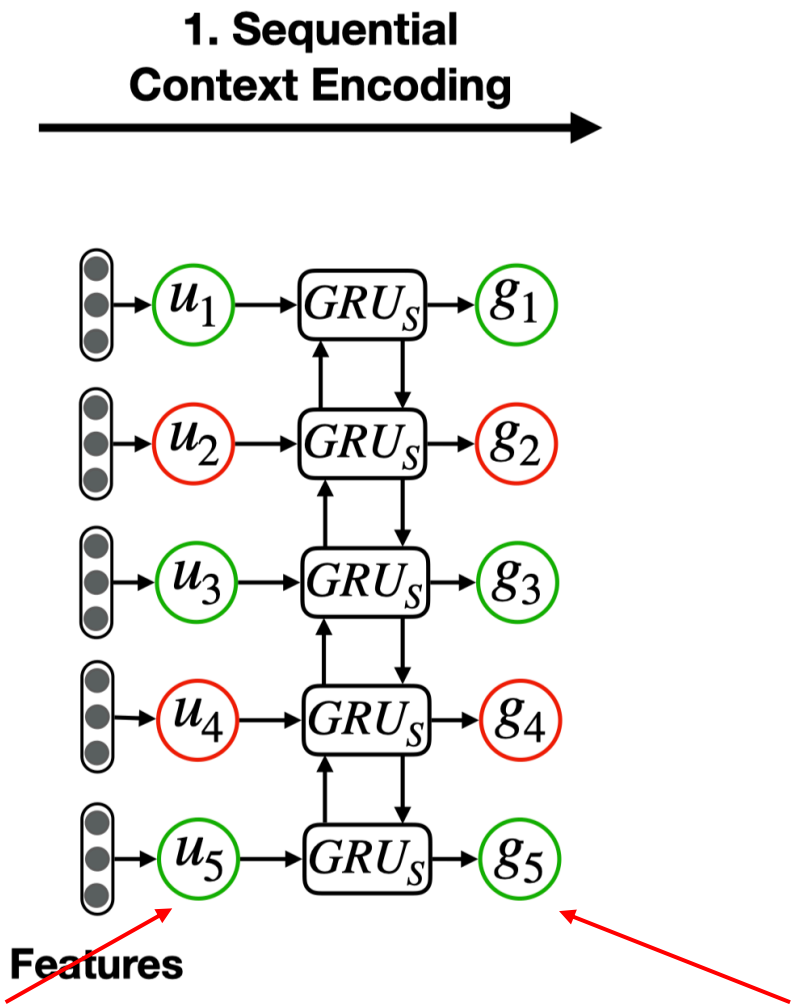
Model

- **Context Independent Utterance-Level Feature Extraction**
- Single convolutional layer followed by max-pooling and a fully connected layer
- This network is trained at utterance level with the emotion labels.

DialogueGCN



Sequential Context Encoder



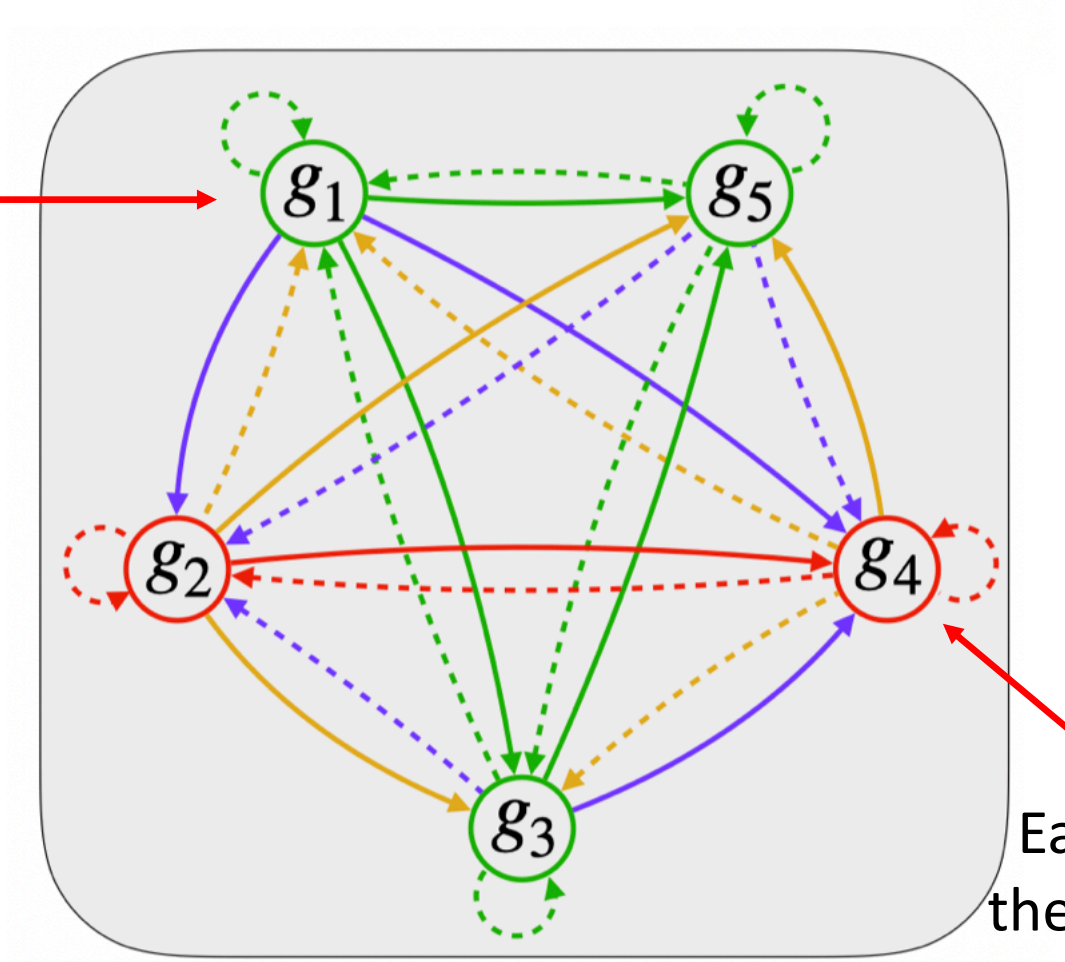
Note : speaker agnostic

context-independent

sequential context-aware

Speaker-Level Context Encoding : vertex

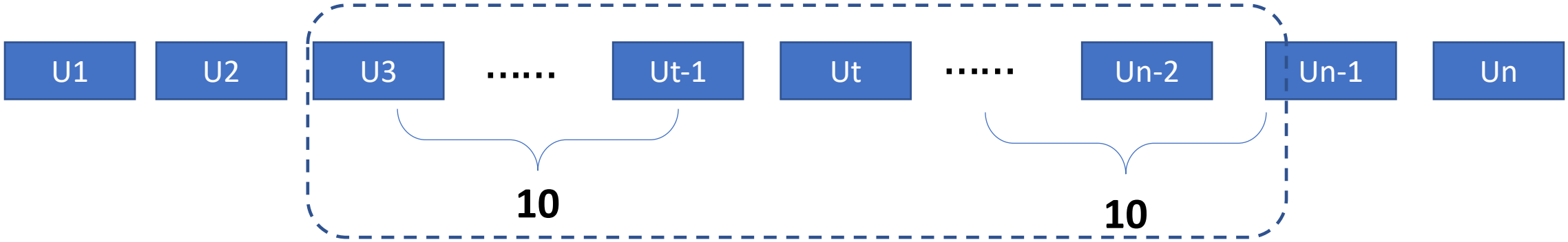
Each utterance in the conversation is represented as a vertex



Each vertex is initialized with the corresponding sequentially encoded feature vector

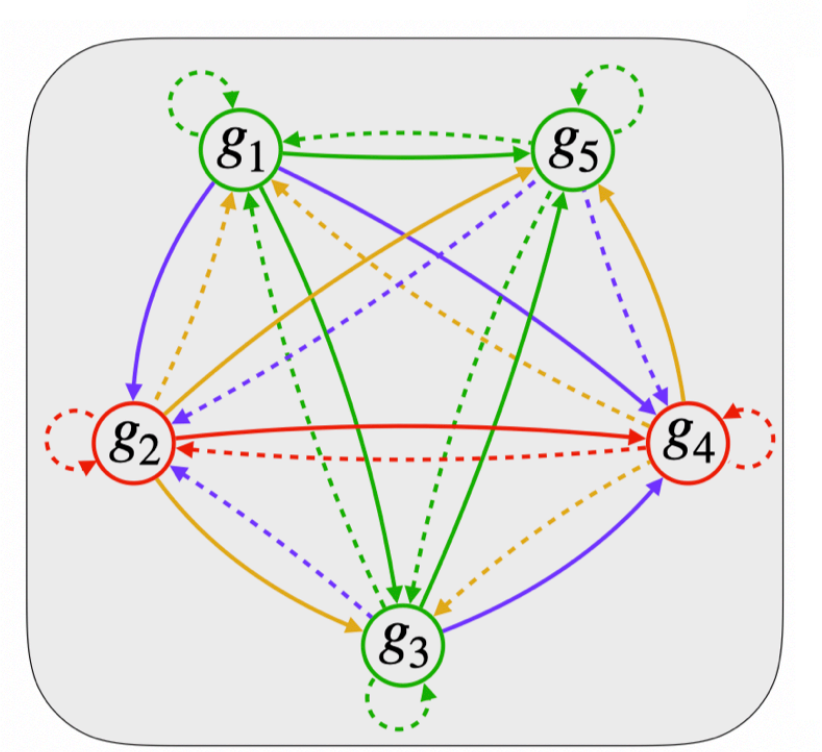
Speaker-Level Context Encoding : edge

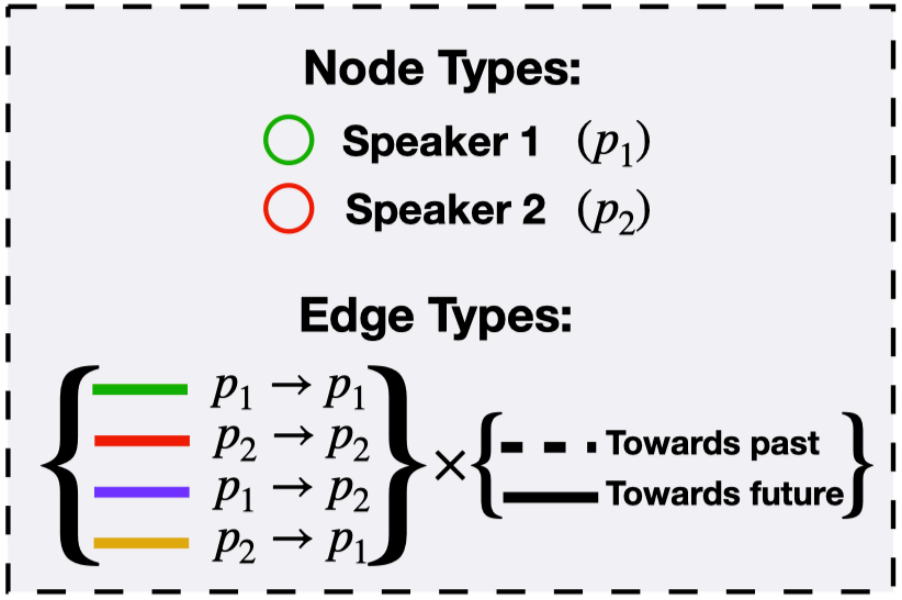
- Keeping a past context window size of *p* and a future context window size of *f*. (=10)



Speaker-Level Context Encoding : edge

- Graph is directed, two vertices can have edges in both directions with different relations
- Relations:





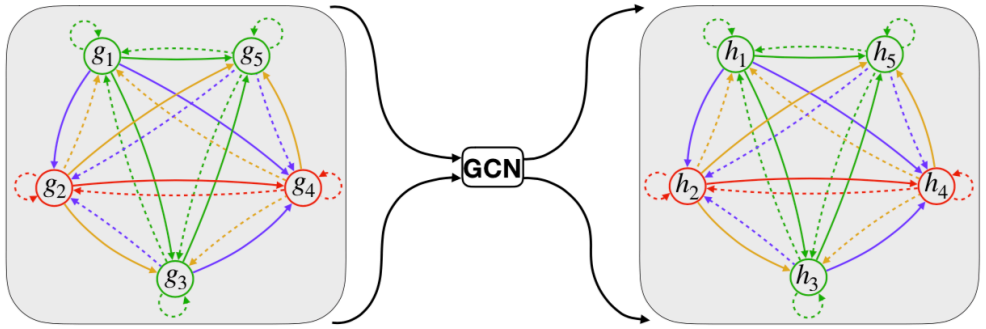
Speaker-Level Context Encoding : transformation

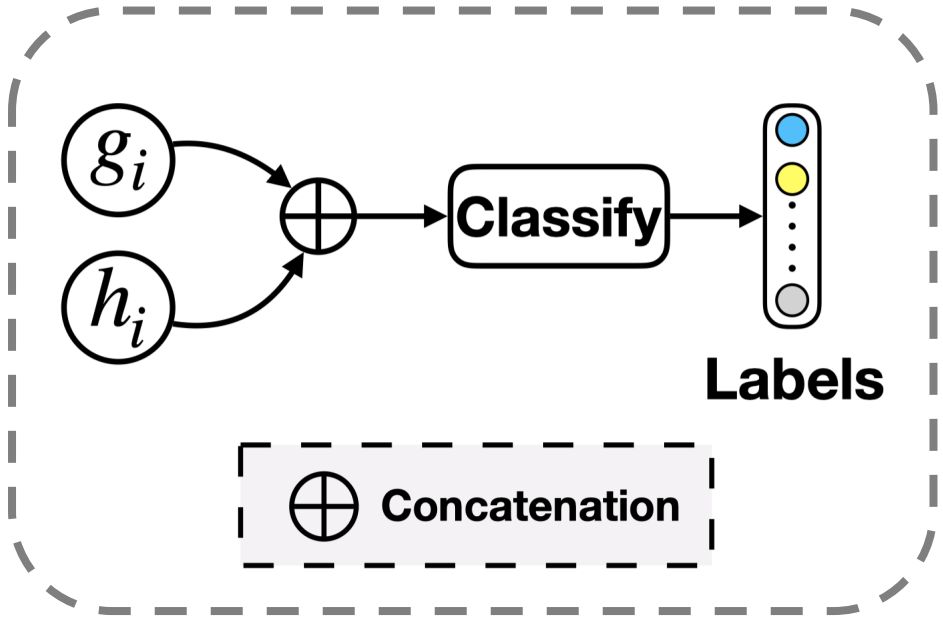
$$\alpha_{ij} = \text{softmax}(g_i^T W_e [g_{i-p}, \ldots, g_{i+f}]),$$
$$\text{for } j = i-p, \ldots, i+f.$$

$$h_i^{(1)} = \sigma\Big(\sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r^{(1)} g_j + \alpha_{ii} W_0^{(1)} g_i\Big),$$
$$\text{for } i = 1, 2, \ldots, N,$$

$$h_i^{(2)} = \sigma\Big(\sum_{j \in N_i^r} W^{(2)} h_j^{(1)} + W_0^{(2)} h_i^{(1)}\Big),$$
$$\text{for } i = 1, 2, \ldots, N,$$

Classification





number of samples/dialogues

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}[y_{i,j}] + \lambda \|\theta\|_2$$

L2-regularization

number of utterances in sample

Dataset

- **IEMOCAP** : happy, sad, neutral, angry, excited, and frustrated.
- **AVEC** : valence ([−1,1]), arousal ([−1,1]), expectancy ([−1,1]), and power ([0,∞)).
- **MELD** : anger, disgust, sadness, joy, surprise, fear or neutral.

Result

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	**65.28**	80.27	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN	40.62	**42.75**	89.14	**84.54**	61.92	**63.54**	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	**64.18**

Table 3: Comparison with the baseline methods on IEMOCAP dataset; Acc. = Accuracy; bold font denotes the best performances. Average(w) = Weighted average.

Methods	AVEC				MELD
	Valence	Arousal	Expectancy	Power	
CNN	0.545	0.542	0.605	8.71	55.02
Memnet	0.202	0.211	0.216	8.97	-
bc-LSTM	0.194	0.212	0.201	8.90	56.44
bc-LSTM+Att	0.189	0.213	0.190	8.67	56.70
CMN	0.192	0.213	0.195	8.74	-
ICON	0.180	0.190	0.180	8.45	-
DialogueRNN	0.168	0.165	0.175	7.90	57.03
DialogueGCN	**0.157**	**0.161**	**0.168**	**7.68**	**58.10**

Result-MELD

1. Multiparty conversations
2. Utterances in MELD are much shorter and rarely contain emotion specific expressions, which means emotion modelling is highly context dependent.
3. The average conversation length is 10 utterances, with many conversations having more than 5 participants.

- Result : new state-of-the-art F1 score of 58.10% outperforming DialogueRNN by more than 1%.

Result-Ablation

Sequential Encoder	Speaker-Level Encoder	F1
✓	✓	64.18
✓	✗	55.30
✗	✓	56.71
✗	✗	36.75

Table 5: Ablation results w.r.t the contextual encoder modules on IEMOCAP dataset.

Result-Ablation

Speaker Dependency Edges	Temporal Dependency Edges	F1
✓	✓	64.18
✓	✗	62.52
✗	✓	61.03
✗	✗	60.11

Table 6: Ablation results w.r.t the edge relations in speaker-level encoder module on IEMOCAP dataset.

Result-Performance on Short Utterances

Emotion of short utterances, like “okay”, “yeah”, depends on the context it appears in.

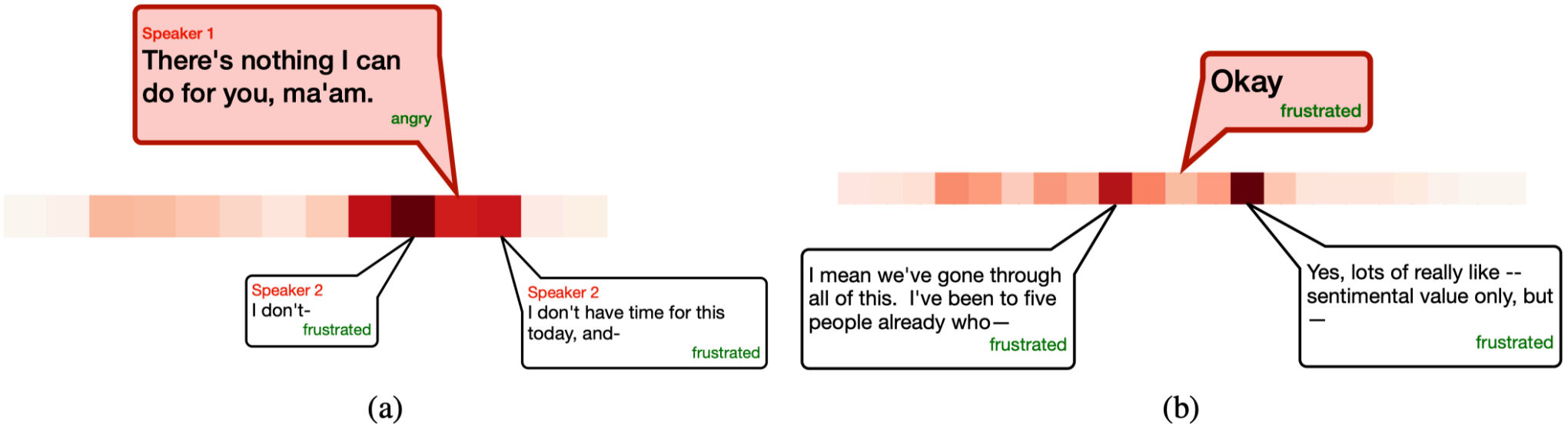


Figure 4: Visualization of edge-weights in Eq. (1) — (a) Target utterance attends to other speaker’s utterance for correct context; (b) Short utterance attends to appropriate contextual utterances to be classified correctly.

Result-Error Analysis

- Frustrated --> angry and neutral
- Excited samples as happy and neutral
  - [subtle difference between two emotions]

- Ok. yes carrying non-neutral emotions were misclassified as we do not utilize audio and visual modality in our experiments.

Thanks!