COIN: Conversational Interactive Networks for Emotion Recognition in Conversation

Haidong Zhang* and Yekun Chai*
Institute of Automation, Chinese Academy of Sciences, China

Abstract

- Existing methods in emotion recognition in conversation tend to overlook the immediate mutual interaction between different speakers in the speaker-utterance level, or apply single speaker-agnostic RNN for utterances from different speakers.
- We propose a conversational interactive model (COIN) by applying state mutual interaction within history contexts and introducing a stacked global interaction module to capture the inter-dependency representation in a hierarchical manner.
- To improve the robustness and generalization during training, we generate adversarial examples by applying the minor perturbations on multimodal feature inputs, unveiling the benefits of adversarial examples for emotion detection.
- The proposed model empirically achieves the current state-of-the-art results on the IEMOCAP benchmark dataset.

Introduction

Task Definition

Emotion recognition in conversation (ERC) aims to detect the speaker's emotions and sentiments within the context of human conversations.

Let there be M parties or speakers $\{p_1, p_2, \dots, p_M\}$ in a human conversation. Given the utterances $\{u_1, u_2, \dots u_N\}$ from a conversation where the utterance u_t is from the corresponding speaker $p_s(u_t)$, the task of ERC is to detect the most likely class from emotion category set C. Here s represents the mapping between the utterances and users.

Challenges

Existing methods mainly includes two categories: one is **modeling each speaker with one RNN** [1]; the other is speaker-agnostic, i.e., **modeling each utterance using one RNN** irrespective of its speaker [2].

However, there is **no direct dyadic interaction between speaker-specific RNNs** in previous work. Different RNNs corresponding to different speakers have been used without mutual interaction [1] or interacting through a mediate global RNN [2]。

References

[1] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. NAACL.

[2] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. AAAI.

* Equal Contribution

Methodology

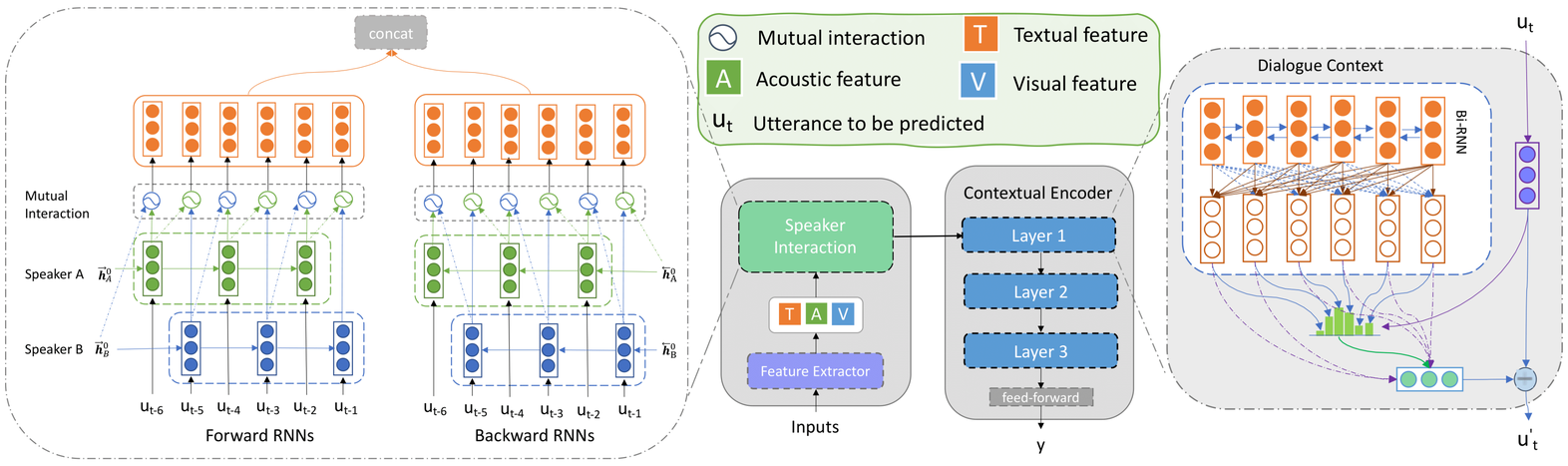

Fig 1. Schematic illustration of COIN

- **Multimodal Features** (including textual, acoustic and visual features) are concatenated along the feature dimension as u_P^i, where $P \in \{A, B\}$ denotes a speaker (here we set two parties or users in the conversation).
- **Mutual Interaction** captures the utterance-level speaker dialogue context with GRUs in two directions, and calculates the interactive representations as:

$$\vec{h}_P^i = \overrightarrow{GRU_P^i}(u_i) \qquad \vec{m}_i = \begin{cases} \vec{h}_A^i \sigma\left(\vec{h}_B^{i-1}\vec{W}_B + \vec{b}_B\right), if\ P = A \\ \vec{h}_B^i \sigma\left(\vec{h}_A^{i-1}\vec{W}_A + \vec{b}_A\right), if\ P = B \end{cases}$$

where h_p^0 represents the initial hidden state of speaker P. The output of both forward and backward direction at step i are concatenated along the feature dimension, denoted as $\overleftrightarrow{m}_i = [\vec{m}_i, \overleftarrow{m}_i]$.

- **Contextual Encoder** consists of $\boldsymbol{L}$ identical stacks, in which of them history dialogue representations M^l is fed into a bi-GRU followed by a self-attention layer to capture the inter-dependency semantics. We can get the context vector for history dialogues:

$$Q, K, V = M_g^l W_Q, M_g^l W_K, M_g^l W_V$$

$$M_{att}^l = softmax\left(d^{-\frac{1}{2}} Q K^T\right) V$$

$$c^l = M_{att}^l softmax\left(M_{att}^l u_t^l\right)$$

$$u_t^{l+1} = \tanh(u_t^l + c^l)$$

- **Adversarial Training** generates adversarial examples by adding perturbations on extracted multimodal features to improve the model generalization.

$$\mathcal{L} = \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{u}) + \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{u_{adv}})$$

where $\mathcal{L}$ refers to cross entropy loss between golden and predicted labels, $u_{adv} = u + \varepsilon \frac{g}{||g||_2}, g = \nabla \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{u})$.

Experiment

Model	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN (Kim, 2014)	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 (Sukhbaatar et al., 2015)	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 (Poria et al., 2017)	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 (Poria et al., 2017)	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	56.91
CMN (Hazarika et al., 2018b)	25.7	32.6	66.5	72.9	53.9	56.2	67.6	64.6	69.9	67.9	71.7	63.1	61.9	61.4
ICON (Hazarika et al., 2018a)	23.6	32.8	70.6	74.4	59.9	60.6	68.2	68.2	72.2	68.4	**71.9**	66.2	64.0	63.5
DialogueRNN (Majumder et al., 2019)	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 (Ghosal et al., 2019)	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
IterativeERC (Lu et al., 2020)	-	53.17	-	77.19	-	61.31	-	61.45	-	69.23	-	60.92	-	64.37
COIN	**53.12**	42.50	85.71	73.07	60.05	62.23	66.48	**68.75**	69.13	69.01	61.73	**66.99**	**66.05**	**65.37**

Table 1: Overall performance of emotion recognition models on IEMOCAP dataset

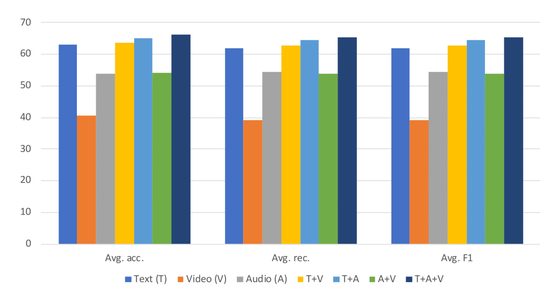

Fig 2. Modality contribution

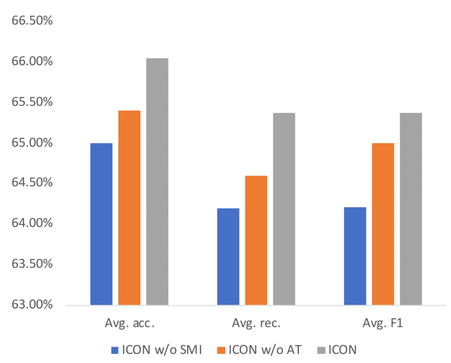

Fig 3. Ablation Study

Conclusion

- We introduce state mutual interaction components to allow for the immediate state interaction between different speakers, and global stacked interaction to capture the contextual and inter-dependency representations.
- We unveil the importance of adversarial training in ERC by promoting the model performance with generated adversarial examples on extracted multimodal embeddings.
- We propose a competing model that achieves the state-of-the-art (SOTA) performance on the IEMOCAP dataset, showing that textual and audio features play the most important role in ERC.