AN INTERACTION-AWARE ATTENTION NETWORK FOR SPEECH EMOTION RECOGNITION IN SPOKEN DIALOGS

NTHU

biic LAB

Sung-Lin Yeh, Yun-Shao Lin, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

Introduction

- A novel attention-based GRU architecture that recognize emotions by taking transactional information into account.
- Our proposed framework extends beyond the conventional framework that often relies on single utterance modeling:
  1) Utilize attention mechanism to embed the transactional information into current utterance representation.
  2) Capture the affective transition from the target speaker and affective influence from the interlocutor to better characterize a target speaker's current emotion state.

Methodology

■ **Dataset:** IEMOCAP Database

- A benchmark dataset that is widely used in speech emotion recognition.
- 10 speakers, 5 sessions, consists of multiple conversational scenarios between two actors.
- Label: Anger, Happiness, Neutrality, Sadness

■ **Feature:** Pitch, Intensity, MFCC (Δ, ΔΔ)

■ **Transactional Context:**

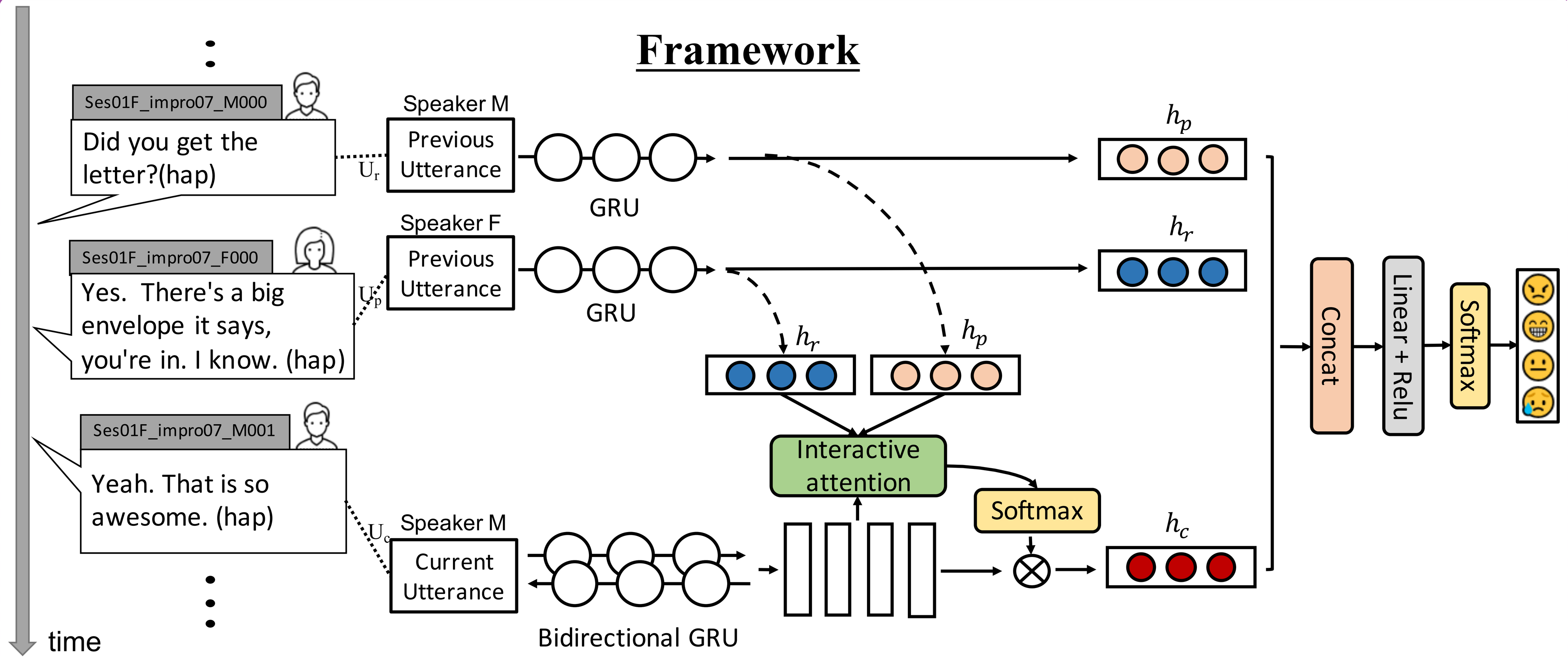
- Previous utterance of the current speaker U_p & previous utterance of the other speaker U_r.
- Each training data point is defined includes a triple of (U_c, U_p, U_r) with the label of U_c.

■ **Interaction-aware Attention (IAA):**

- *Score function*: $e(h_{it}, h_p, h_r) = v^T \tanh(W_c h_{it} + W_p h_p + W_r h_r + b_a)$
- *Attentive weight*: $\alpha_t = \frac{\exp(e(h_{it}, h_p, h_r))}{\sum_{t=1}^{T} \exp(e(h_{it}, h_p, h_r))}$
- *Context vector*: $h_c = \sum_{t=1}^{T} \alpha_t h_{it}$

Experimental setup and results

Framework



Interactive attention : The defined score function iteratively loops through every timestep of current utterance's hidden states h_{it} based on the contextual representations of h_p and h_r.

Model	Method	Recall(%)				WA(%)	UA(%)
		Anger	Happiness	Neutrality	Sadness		
SVM Trees	Rozgić et al. [1]	-	-	-	-	60.8	60.9
BiLSTM+ATT	Mirsamadi et al. [2]	-	-	-	-	63.5	58.8
CMN	Hazarika et al. [3]	-	-	-	-	65.3	-
MDNN	Zhot et al. [4]	-	-	-	-	61.8	62.7
BiGRU+ATT	Our method	56.6	59.4	48.4	71.6	57.6	58.4
BiGRU+IAA	Our method	65.3	61.0	51.7	73.0	60.7	62.9
RandIAAN	Our method	66.0	62.3	53.5	73.7	62.0	63.4
IAAN	Proposed method	72.1	65.4	53.1	74.6	64.7	**66.3**

Model Variants

BiGRU+ATT: A BiGRU network with the classical attention (ATT) trained only using current utterances.

BiGRU+IAA: A BiGRU network with IAA, but the final prediction only depends on current utterance's representation.

RandIAAN: IAAN trained with randomly selected utterances in the dialog as a transactional frame.

Analysis of Different Transactional Contexts

Scenario	Data points(%)	BiGRU+ATT		IAAN	
		WA(%)	UA(%)	WA(%)	UA(%)
Case 1	27.1	72.3	69.2	76.0	74.4
Case 2	47.1	58.2	60.2	64.3	66.6
Case 3	25.8	42.0	42.4	53.6	54.8

Case 1: U_c shares the same emotion as U_p and U_r.

Case 2: U_c shares the same emotion with one of U_p and U_r.

Case 3: U_c has emotion different from one of U_p and U_r.

Conclusion

- Our interaction-aware attention allows more compact current utterance representation compared with classical attention mechanism.
- The contextual information is effectively incorporated both at current utterances representations learning and final prediction stage in dyadic conversations.
- Our method shows outstanding performance with unweighted accuracy of 66.3%, and outperforms the best known traditional and state-of-the-art methods.

Future Work

- Validate the robustness and generality of our IAAN in other conversational dataset.
- We observe that transactional information can be misleading; thus, developing a strategy that is able to consider the strength of influence from emotional contexts is of importance.

References

[1] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on . IEEE, 2017, pp. 2227–2231.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[3] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P.Morency, and R.Zimmermann, "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) , vol. 1, 2018, pp. 2122–2132.