A Self-Attentive Emotion Recognition Network

Cyprus University of Technology

Harris Partaourides, Kostantinos Papadamou, Nicolas Kourtellis, Ilias Leontiadis, Sotirios Chatzis

{c.partaourides, sotirios.chatzis}@cut.ac.cy, ck.papadamou@edu.cut.ac.cy
{nicolas.kourtellis, ilias.leontiadis}@telefonica.com

ENCASE ENhancing seCurity and privAcy in the Social wEb

Emotion Recognition in Online Text Chats

- Recognizing emotions is an ambiguous process with high dependence on the contextual information both on the utterance and dialog level.
- State-of-the-art approaches exploit whole dialog knowledge and have difficulties capturing temporal dependencies over long horizons.
- Accurate real time emotion recognition is paramount importance for early identification of cyberbullying and suicidal ideation in Online Social Networks (OSNs).

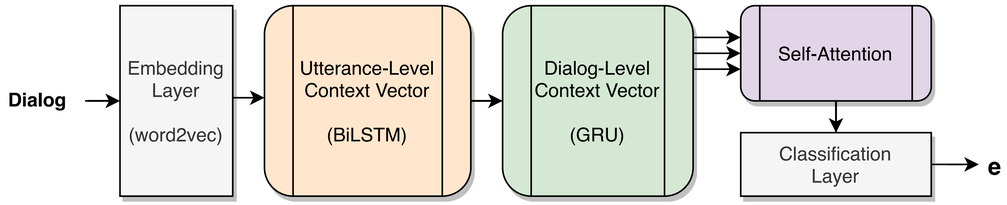
Dataset

- We utilize a well-known benchmark for emotion recognition, namely the IEMOCAP dataset.
- The dataset has been collected by emulating conversations in a controlled environment in order to study expressive human behaviors.
- The conversations have been performed by ten unique speakers over five dyadic sessions in both a scripted but also an improvisation manner with various audio-visual modalities being recorded.
- Each utterance in the dataset is labeled by three human annotators using categorical labels; these include angry, sad, happy, frustrated, excited, neutral as well as other categories which we omit in this study.
- The available annotation has been performed by three annotators who assess the emotional states of the speakers taking into consideration dialog context.
- We only utilize the textual modality (prevalent form in OSNs) and the label information derived by performing majority voting.
- The dataset contains 151 conversations with a total number of 10,039 utterances. However, only 7,380 utterances contain the six types of emotions we retain in this study.

Self-Attentive Emotion Recognition Network

For the first time in the literature, we introduce a *self-attentive hierarchical encoder* network that is capable of extracting salient information on both the individual utterance level as well as the level of the dialog context, as it has evolved until any given time point.

- A trainable *Word2Vec embeddings* mechanism is presented with the input sequence.
- A *bidirectional Long Short-Term Memory* (BiLSTM) is used to capture the salient lingual information contained within each utterance.
- A *Gated Recurrent Unit* (GRU) that performs dialog context-level representation to allow for capturing the salient dynamics over the whole dialog span.
- A *self-attention layer* on top of the dialog-level GRU network.



- The devised model is trained in an end-to-end fashion.
- The employed training objective function is the categorical cross-entropy of the model.

Utterance at t	t-10	t-9	t-8	t-7	t-6	t-5	t-4	t-3	t-2	t-1	t	Emotion
Then she's gone.	0.02	0.45	0.02	0.02	0.03	0.32	0.02	0.08	0.00	0.00	0.04	Sad
It's going to be ...	0.29	0.00	0.00	0.01	0.39	0.00	0.02	0.00	0.00	0.00	0.28	Sad
Well, you know ...	0.04	0.03	0.04	0.32	0.03	0.10	0.01	0.01	0.03	0.33	0.07	Sad
Sure	0.05	0.08	0.21	0.08	0.11	0.07	0.01	0.07	0.11	0.09	0.12	Sad
to talk to somebody ...	0.01	0.51	0.00	0.03	0.00	0.00	0.00	0.30	0.00	0.00	0.14	Sad
you shouldn't be ...	0.32	0.00	0.01	0.00	0.00	0.00	0.10	0.00	0.00	0.02	0.55	Sad
It's just going to ...	0.00	0.01	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.37	0.54	Sad
Yes.	0.02	0.02	0.00	0.01	0.05	0.02	0.01	0.04	0.22	0.43	0.17	Sad
Ah.	0.00	0.00	0.00	0.13	0.00	0.00	0.04	0.41	0.33	0.00	0.08	Sad
Thank you.	0.00	0.01	0.06	0.03	0.02	0.04	0.16	0.29	0.18	0.09	0.10	Sad

Experimental Evaluation

	Accuracy	Precision	Recall	F1 Score
SVM	0.313 (±0.00)	0.484 (±0.00)	0.235 (±0.00)	0.316 (±0.00)
$BiLSTM$	0.477 (±0.01)	0.471 (±0.02)	0.459 (±0.01)	0.465 (±0.01)
$BiLSTM_{att}$	0.516 (±0.02)	0.516 (±0.02)	0.501 (±0.02)	0.509 (±0.02)
SERN	0.522 (±0.02)	0.544 (±0.02)	0.517 (±0.02)	0.530 (±0.02)

	Angry	Excited	Frustrated	Happy	Neutral	Sad	Recall
Angry	110	2	29	0	22	7	0.647
Excited	9	156	8	74	27	25	0.522
Frustrated	71	6	193	1	87	23	0.507
Happy	14	19	0	80	29	1	0.559
Neutral	35	34	83	11	197	24	0.513
Sad	9	12	42	7	11	164	0.669
Precision	0.444	0.681	0.544	0.462	0.528	0.672	

Classes	Accuracy	Precision	Recall	F1 Score
4	0.689 (±0.03)	0.685 (±0.02)	0.699 (±0.02)	0.692 (±0.02)
5	0.583 (±0.02)	0.589 (±0.02)	0.569 (±0.02)	0.579 (±0.02)
6	0.522 (±0.02)	0.544 (±0.02)	0.517 (±0.02)	0.530 (±0.02)

	Accuracy	Precision	Recall	F1 Score
$SERN_5$	0.557	0.563	0.552	0.558
$SERN_{10}$	0.570	0.570	0.591	0.581
$SERN_{20}$	0.584	0.583	0.580	0.582
$SERN_{40}$	0.581	0.595	0.565	0.579
$SERN$	0.555	0.555	0.570	0.562

Utterance at t	t-10	t-9	t-8	t-7	t-6	t-5	t-4	t-3	t-2	t-1	t	Emotion
Oh, you infuriate me ...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	Angry
Yeah, well I ignore ...	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.06	0.20	0.70	Frustrated
And she-she's Larry's ...	0.00	0.00	0.00	0.02	0.00	0.04	0.24	0.01	0.00	0.00	0.69	Frustrated
Well, from your father's ...	0.00	0.00	0.01	0.00	0.00	0.16	0.05	0.02	0.00	0.07	0.68	Frustrated
Cause listen, I'm telling ...	0.01	0.07	0.00	0.03	0.15	0.02	0.00	0.01	0.20	0.28	0.21	Frustrated
What do you want from ...	0.07	0.01	0.00	0.05	0.09	0.13	0.03	0.02	0.22	0.08	0.30	Frustrated
Every time I reach out ...	0.00	0.02	0.12	0.00	0.00	0.00	0.11	0.13	0.07	0.02	0.54	Frustrated
You're a considerate ...	0.02	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.92	Neutral
To hell with that.	0.01	0.10	0.69	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.15	Angry

Acknowledgments

This research is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025).