An Enhanced Speech Emotion Recognition System using Attention Network

Kumar T. Rajamani[1], Srividya Rajamani[2]

[1] Institute of Medical Informatics, University of Luebeck

[2] University of Augsburg

kumar.rajamani@uni-luebeck.de
kumartr@gmail.com

Abstract:

Speech emotion recognition (SER) is a challenging area of study where a lot of research has been dedicated in recent years to boost the performance. It is an integral component and is key for building human-machine interface. The initial research efforts have largely considered each utterance of the speaker and the interlocutor in isolation. There are transactional dependencies between human-human conversation and this was further captured in the recent work on Interaction-Aware Attention Network (IAAN) by Yeh et al (1). IAAN incorporates contextual information by using attention mechanism. The dataset where this was demonstrated is the IEMOCAP dataset, which is a benchmark dataset used widely in speech emotion recognition which has a total of 5531 utterances. The baseline achieved by IAAN is 66.3% unweighted accuracy (UA) and weighted accuracy (WA)of 64.7% in the four class emotion (anger, happiness, sadness and neutrality) recognition challenge. We have utilised the implementation provided by Yeh at al (1) in our work and explored boosting the accuracy of performance of the proposed network. The number of epochs in the initial implementation was limited at 3000 epochs. Gated Recurrence Units take longer to train, and hence we explored increasing the number of epochs to 6000. Our experiments with increased number of epochs gave us substantial gains in the performance of SER, by obtaining 68.05% unweighted accuracy (UA) and weighted accuracy (WA) of 66.44%. We have achieved 1.75% gain in unweighted accuracy (UA) and 1.74% gain in weighted accuracy (WA) and is the current state-of-art recognition rates obtained on the benchmark database. The accuracy numbers do vary slightly from one execution to the next and we propose in our next steps to average the accuracies over multiple runs to obtain reliable and repeatable results.

References

1) S. Yeh, Y. Lin, and C. Lee, "An interaction-aware atten-tion network for speech emotion recognition in spokendialogs," inICASSP 2019 - 2019 IEEE InternationalConference on Acoustics, Speech and Signal Processing(ICASSP), 2019.