Multimodal Integration for Empathy Recognition

Nikoletta Xirakia[1], Tayfun Alpay[1], Pablo Barros[2], Stefan Wermter[1]

[1] University of Hamburg - {6xirakia, alpay, wermter}@informatik.uni-hamburg.de
[2] Italian Institute of Technology - pablo.alvesdebarros@iit.it

Empathy is one key ingredient in human-human interactions that enhances our communication and understanding of each other and improves our interpersonal relationships. Considering the rapid pace within which social agents have been entering our lives over the past years, the need to improve our social interactions with them becomes more urgent. Agents that can predict human affective behavior can result in more natural human-robot interactions.

This work proposes a model that integrates data from multiple modalities, audio, vision, and language, to recognize valence-based affective behavior in a dyadic interaction setup. We investigate the complexity of integrating and processing asynchronous multi-sensori data by exploring the capabilities of two novel recurrent neural network architectures, the SkipRNN and Phased LSTM, designed to ignore irrelevant and repetitive input data. We compare the performance of the two architectures to the conventional LSTM. We also examine the eGeMAPS acoustic set's performance, designed to be applied to valence-based affective speech recognition tasks, by comparing it against the state-of-the-art MFCC acoustic set. Our model is trained and evaluated on the OMG-Empathy datasets that employ valence-based annotations, and it is set around semi-scripted dyadic human interactions.

Our results show that valence-based tasks require higher volume datasets to properly generalize the task, where most of our experiments suffered from overfitting the training set. In terms of synchronizing the different modalities, the SkipRNN achieved the highest performance across all other experiments, where the baseline LSTM achieved a comparable performance. In contrast, Phased LSTM achieved the worst performance across all experiments, which we believe is due to the small size of the OMG-Empathy dataset, combined with customizing the network to ignore any repetition in the input data.

Finally, the acoustic set performance results showed that eGeMAPS could increase the valence recognition accuracy when combined with the LSTM model. At the same time, it achieved a significantly worse performance in the SkipRNN and Phased LSTM models compared to the MFCC acoustic set. This demonstrates that eGeMAPS require high volume valence-based datasets to increase the learning performance in tasks that aim to recognize affective behavior.