John Kyle Cooper
Dr. Duan
ECE 477
12/06/19

Attention Based 3-D Convolution Neural Network for Speech Emotion Recognition

Author: Meiying Chen

Summary of the paper.

The proposed study uses a 3-D Convolutional Neural Network (CNN) to process audio-visual data of human speech/conversation and output the emotion conveyed from inputted speech. The proposed method (3D CNN) outperforms the baseline models: a CNN model by S. Poria et al. (2016) and a convolutional recurrent neural network (CRNN) model by Y. Zhao et al. (2017).

Good things about the paper.

The introduction of this paper helped me understand the importance and utility of machine (i.e. machine-learning model) recognition of human emotions. Furthermore, the introduction helped me understand the problems the current researchers in the field of speech signal processing are facing with regard to different aspects of using machine learning models to recognize emotions. The methods section is explained well Also, the author performed a thorough and impressive literature review.

Major comments.

In the introduction, the author states that methods like Support Vector Machines (SVMs) have difficulty processing speech in other languages for emotion recognition, but the author does not state the reason for this difficulty. I suggest adding a few sentences stating the reason SVMs have difficulty recognizing emotions in other languages (e.g. are the SVM models only trained on English language data?).

In section 3.2, I did not understand what "textual modalities" are since they are not defined by the author in this section (e.g. are textual modalities considered features such as closed captioning?). I suggest providing a short sentence describing what these textual modalities are with respect to the input data.

The Results and Conclusion sections were not written yet and there are some figures and tables that have not yet been made, which is understandable since this is just a draft of the project report. However, I would have liked to give you feedback on those sections (I'm glad the author reports that the model is better than the baseline model and is much faster. I am looking forward to seeing their final project!).

Minor comments.

- In section 2.1, the *d* in the delta features (defined by m_i^d) is not specified.
- In section 2.3, *W*, *b*, and λ variables are not defined by the author.
- There are not that many grammatical corrections that need to be made before the final paper. Therefore, I will not send a scanned copy of the project report to the author. However, if the author wishes to reach out to me once they have their final draft prepared, I would be happy to proofread it and provide more feedback.

Reference:

1. *Meiying Chen (2019). Attention Based 3-D Convolution Neural Network for Speech Emotion Recognition.*