ATTENTION BASED 3-D CONVOLUTIONAL NEURAL NETWORK FOR SPEECH EMOTION RECOGNITION

Author: Meiying Chen

Summary:

In this paper the author tackles the problem of emotional recognition from speeches. They use the IEMOCAP dataset to train and test the model. The author uses the log-Mels to transform the speeches and then as inputs to the model. The author proposes a model consisting of a 3D CNN that contains 7 layers and a linear layer at the end. They also try a novel approach by incorporating Attention Layer after the linear layer to extract important features related to emotions from the last layer of convolution. Finally, this performs better than the baseline models.

Strengths

I like how the author used attentions models to extract important features. It makes sense as certain emotions must have distinct features that define them and so focusing on them should improve classification. I really liked how the paper also had a well-researched background section.

Major Comments:

It seems to me that the dataset contains speeches from two individuals only. I have doubts on whether that makes the model generalizable. The model will probably overfit on the emotions for the two individuals.

This isn't a criticism but rather an improvement that I am proposing. Attention Models could be used find what aspects of the log mel spectrogram that could make the model more interpretable. For instance, maybe some emotions have a distinct power and frequency combination. I say this as there has been research done in the past that uses attention models to understand how the model is learning.

Minor Comments

- Description on the dataset of whether the emotional speeches were posed or unposed (if so, were they hand labelled)
- Average length of the recordings.