Introduction and Significance

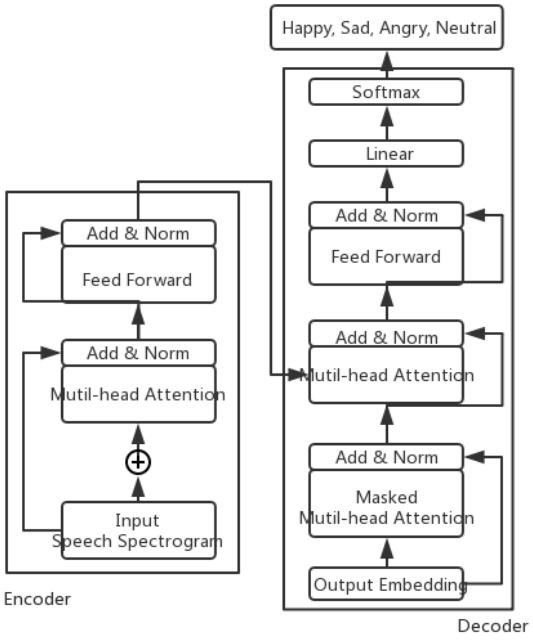
Emotions play a crucial role in human communications and successfully detecting the emotion states is helpful to improve the efficiency of human-computer interaction. There has been a number of previous works on speech emotion recognition with machine learning and deep learning methods. However, most of them ignore the sparsity of human emotions in the audio pieces. That means, in most cases, human emotions can only be detected in some specific moments during a long utterance, or say, not every frame in a speech recording contains emotional information. If this sparse characteristic can be utilized so that different weights are adaptively assigned to different locations according to the relevance of positions and emotions in a signal, this kind of integrated model may be more accurate and efficient.

Unfortunately, existing models lack an efficient way of writing information to a long-term memory component and utilize these memories in the predicting process. Some former studies have applied recurrent neural networks like LSTM/BLSTM architecture to employ the previous information of each signal location i.e. memories. Nevertheless, this kind of memory information exists in the form of hidden layer and neuron weights, which is not sufficient for predicting, as what is the most relevant information that passed between the past and future remains unknown.

Attention models fixed this problem and avoid emphasis too much on the data point being close to one another by allowing the model to automatically search for parts of the source that are relevant to predicting the target, without fixing the length of the vectors used. Therefore, we propose a kind of encoder-decoder model based one of the famous attention architecture Transformer, to further explore the potential of long-distance memories.

Apart from the sparsity problem, there is other ignorance of the current models. For example, the emotional states in a video or speech may be different at different times. So, it is better to consider emotion recognition as a detection task which detects changes in emotional states over each moment, rather than simply view it as a classification work of the whole audio piece. The proposed model also has the potential to tackle this problem as it can be modified to generate predicted emotion sequence.

Model Design



Plans

Intended dataset: IEMOCAP

Baseline model: SVM, BLSTM-RNN

Time Plan: totally 8 weeks

week1-2: literature review, dataset search

week3-5: implement proposed models

week 6-7: write paper and poster

Task allocate plan:

All work allocated to I, me and myself