Project Summary
of
Speech Emotion Recognition

Group Name: Lucky Tiger

z5382241 Jinlei Ru

z5373485 Vivien Xian

z5212886 Zhiqing Cen

z5212992 Zhaocheng Li

Introduction

Speech is the expression of ideas and emotion through sound. Due to the nature that speech carries emotion, it is a medium for the listeners to identify feelings and attitude of the speaker. It plays a vital role in daily communication among humans. As the demand of human-computer interaction increases, Speech Emotion Recognition (SER) is commonly used to extract emotional state through processed and classified speech signals. However, correctly recognising human emotion from speech is a complex and challenging task since emotions are subjective.

This report aims to detect and identify substantive emotion in voice recording using deep Learning.

Literrature Review

Chen et al. (2018) proposed to use 3-dimensional attention-based convolutional recurrent neural networks (ACRNN) with Mel-spectrogram for speech emotional recognition. Four emotions classifications are used, angry, sad, happy and neutral, with an accuracy of 64.74% on the IEMOCAP dataset. With 70.47% accuracy on angry, 84.32% on sad and 66.52% on neutral emotion, interestingly, there is only a 29.95% accuracy on happy emotion.

Methods and Models

Extracting target emotion labels

The emotion labels of those sentence audios are stored in different txt files, for easier processing in later steps, and to avoid repeated work, we extracted those labels from text using regular expression and stored the file names of all sentence audios together with their emotion labels in a data-frame, which was then written to a csv file.

Extracting input audio data

Since all sentence audios are in wav format, they need to be transformed into data structure that allows easier feature extraction and data analysis. To achieve this, a python package named 'librosa' is used, which provides function for retrieving audio information (by loading audio files as floating-point time series), as well as a wide range of building blocks necessary to pre-process audio, extract various features and visualize those audio features.

In the dataset, a speech was sampled at a rate of 16,000 audio data points per second, and the native sampling rate when loading each audio file is applied. After that, because some audio can be quite long, it has to trim any leading or trailing silence by setting the threshold as 20 decibels (dB). Then, noise reduction to the trimmed waveform is performed, so that there won't be noises degrading the performance of our speech emotion recognition models.

After these pre-processing procedures, the waveform can actually be used directly as input to our MLP and CNN models as long as any shorter ones are padded to the same length. However, it is found that the size of data would be so large, which makes the training of NNs too long while did not provide extra benefit on prediction performance. So, extracted features as input instead of the waveform is used. Two types of features as well as the combination of them are tried, which are Mel spectrogram and Mel-frequency cepstral coefficients (MFCC). For a given waveform, both of Mel spectrogram and MFCCs are originally extracted as matrices, the means of those 2-d arrays is took to transform them into 1-d arrays, which is a standard treatment in audio analysis. To avoid padding, when taking means, arrays should be transposed, which did not reduce performance of our models while providing a fixed size of input. This trick also allowed the models to have less weights to learn, thus reduced running time requirement.

Experimental Setup

In this project, IEMOCAP (The Interactive Emotional Dyadic Motion Capture Database) is used as the dataset for speech emotion recognition. In this dataset, each audio is evaluated by at least three different judgers. At the same time, IEMOCAP uses motion capture, and the expressions of actors as well as their movements are recorded. This helps the judges to accurately judge their emotions. The emotions evaluated include angry, happy, sad, neutral, frustrate, excited, fearful, disgusted, and others.The audios were based on improvisation or scripts. The best 3 scripts are chosen from more than 100 of 10-minute scripts. And there are 5 sessions in this data set.

Before modelling, since there are some categories of emotions not suitable for the speech emotion recognition problems. Filtering out or combining

some of them before splitting our data into training, validation, and testing are needed.For binary classification, happy and neutral are combined into calm, while combined the other emotions including angry, sad, frustration and fear into uncalm.For Multi-class classification, we kept only four categories: angry, sad, neutral, and happy emotions.

Modelling

Two types of Neural Networks are explored. For baseline model, the MLPClassifier in Scikit-learn package is used, which provides convenient interface for implementing a wide variety of fully connected feed-forward MLP models with different architectures and parameters.

A more advanced model we tried was 1-d CNN(built from scratch using Pytorch), which is mostly used in time-series data while also used on audio or other signal data. Unlike MLP, which views each input unit independently and does not capture correlation in input, 1-d CNN is able to extract extra feature from the input array by utilizing kernel that slides along one dimension.

Results

The results in Table below show the train accuracy and test accuracy for the lbfgs optimizer for the initial MLP model using binary approach to separate emotions from the dataset into calm and uncalm.

Model	Method	optimizer	alpha	Hidden layer sizes	Train Accuracy	Test Accuracy	Loss
MLP	Binary {calm, uncalm}	lbfgs	0.13	50,30,10	99.63%	65.71%	0.07%

Based on results in table above, the MLP model did not perform well for test accuracy because our model was over-fitting, therefore, the optimizer should be changed in the MLP model, and the results in Table below show the results.

Model	Method	optimizer	alpha	Hidden layer sizes	Train Accuracy	Test Accuracy	Loss
MLP	Binary {calm, uncalm}	adam	0.13	50,30,10	77.05%	71.43%	0.89%
MLP	Binary {calm, uncalm}	sgd	0.13	50,30,10	83.36%	71.21%	0.95%

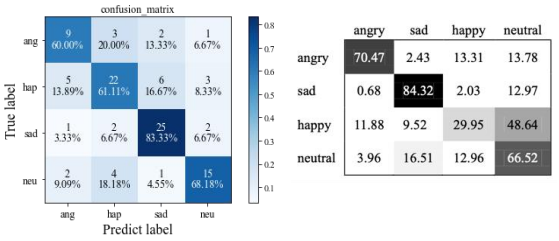
Comparing the results above, test accuracy is increased because adam and sgd optimizer is suitable for large dataset. Then, four categories are used to separate emotions from dataset into angry, happy, sad, and neutral, and the results in table below.

Model	Method	optimizer	alpha	Hidden layer sizes	Train Accuracy	Test Accuracy
MLP	Four categories {angry, happy, sad, neutral}	adam	0.13	50,30,10	35.27%	27.68%
MLP	Four categories {angry, happy, sad, neutral}	sgd	0.13	50,30,10	33.28%	26.54%

Therefore, CNN model is used to analysis speech emotion recognition, and the tables below show that the train accuracy and test accuracy by CNN model by different optimizer. According to the tables, the results show that the extraction method has significant impact on test accuracy, and sgd is best optimizer in CNN model.

Model	Method	optimizer	Data size	extraction method	Train Accuracy	Test Accuracy
CNN	Four categories {angry, happy, sad, neutral}	sgd	All 5 sessions	mfcc & mel	96.34%	65.16%
CNN	Four categories {angry, happy, sad, neutral}	sgd	All 5 sessions	mfcc	96.02%	61.19%
CNN	Four categories {angry, happy, sad, neutral}	sgd	All 5 sessions	mel	96.67%	22.74%
Model	Method	optimizer	Data size	extraction method	Train Accuracy	Test Accuracy
CNN	Four categories {angry, happy, sad, neutral}	adam	All 5 sessions	mfcc & mel	95.32%	63.36%
CNN	Four categories {angry, happy, sad, neutral}	adam	All 5 sessions	mfcc	94.53%	60.83%
CNN	Four categories {angry, happy, sad, neutral}	adam	All 5 sessions	mel	89.58%	25.63%

Finally, we present the confusion matrix to further analyze the SER performances of CNN model. According to Fig.1, 16.67% happy samples are misclassified as sad, 13.89% happy samples are misclassified as angry, and 8.33% happy are misclassified as neutral. Compared to the four categories emotion analysis model of the IEMOCAP database in 3-D Convolutional Recurrent Neural Networks with Attention Model (Fig. 2), our results show a significant improvement.



Figures 1&2, 1: confusion matrix of CNN; 2: confusion matrix of Chen's results

Conclusion

In conclusion, the best result obtained in this report is to use four category labels, angry happy, sad, and neutral. Extracting audio features of these voice recording into MFCC and Mel-spectrogram, then passed into Convolutional Neural Network (CNN) with sgd as the optimizer. With an accuracy of 65.16% similar to the 64.74% in Chen's et al. (2018) paper, this result has a 61.11% accuracy on recognising the happy emotion which is twice higher than 29.95% in his result. And with only 8.55% chance to recognise happy emotion into neutral emotion, it can be said that the limitation of misidentifying happy as neutral emotion in Chen's result has been overcome.