Speech Emotion Recognition in Using Convolutional Neural Network

Project of group 4: (ID: 17060- 46, 47, 48, 49) EEE 312

This is a project on speech emotion recognition using the Ryerson Audio-Visual database of emotional speech and song

Emotion and Motivation

Emotion is a mental state associated with the nervous system in an animal body. As humans we feel emotion every day. These feelings are somewhat personal and subjective, and hence it is not surprising that science have always struggled to grasp the proper understanding of it. Neural networks and computer vision have taken this challenge in recent times to decipher human emotions and have pushed it one step further by predicting, in broad terms, the emotion of a human beings by analyzing tonal properties, body gestures and facial expressions. The motivation behind taking this as our project was to test how well we could train a model to predict something as subjective as emotion from pure audio files and push to attain a respectable level of accuracy.

The dataset

The dataset that we have used for training the model is Ryerson Audio-Visual database of emotional speech and song (RAVDESS). It contains 7356 files. The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Here we decided to make a model that would only use the audio files to predict the emotion. So, we used the audio-only files from this repository. There are two types of audio files here, one is called the emotional speech audio and the other was the emotional song audio. Although initially the emotional speech audio was used, it was felt that our model lacked data to be trained on and hence we added the song audio as well. The emotional speech audio had 1440 audio files with recordings 12 male and 12 female actors and the emotional song audio had 1012 audio files with the recordings of those same actors. The audio files are all approximately 4 seconds long.

The audio only database include these two statements: 01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door" .

The most striking part of this dataset was that the validation accuracy of the preexisting models is lower that what one would expect from the neural networks at the time of writing, and so a brief comparison between our models and the pre-existing models is drawn later.

The Preprocessing

Our goal was to use a one dimensional convolutional neural network to train our model with the RAVDESS dataset. The input of the convolutional layer was a matrix with values referred to the features that were extracted from each of the audio files.

The following features were extracted:

ZCR: zero crossing rate. This is the rate at which out audio file crosses the x axis.

STFT: short time Fourier transform: The STFT represents a signal in the time-frequency domain by computing discrete Fourier transforms (DFT) over short overlapping windows

Chroma stft: Compute a chromagram from a waveform or power spectrogram.

Mfcc: mel-frequency cepstrum. This is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Rms: root mean square.

Mel: computes a mel scaled spectrogram.

Tonnetz: Computes the tonal centroid features (tonnetz). This projects chroma features onto a 6-dimensional basis representing the perfect fifth, minor third, and major third each as two-dimensional coordinates. This was later added since we decided to use the song dataset and tonnetz is known to be an effective feature to be extracted from songs.

The data was also augmented by noise injection and by change in pitch (stretching and compressing). Thus, after augmentation we ended up with three times the number of original data. Original, noise injected and pitch changed.

The data was then divided into training and testing sets. Here, since we are aiming for high accuracies on the RAVDESS in the project, there is no outside data as testing data and hence our validation set is termed as testing set in the code. The validation set was 25% of the original testing set.

The proposed model

The following is the model architecture that we have implemented:

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 168, 256)	1536
max_pooling1d (MaxPooling1D)	(None, 84, 256)	0
conv1d_1 (Conv1D)	(None, 84, 256)	327936
max_pooling1d_1 (MaxPooling1	(None, 42, 256)	0
conv1d_2 (Conv1D)	(None, 42, 128)	163968
max_pooling1d_2 (MaxPooling1	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_3 (MaxPooling1	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 8)	264

Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0

*this model is inspired from one of a Kaggle user Shivam Burnwal | Expert | Kaggle. However some parameters were changed in the model where it was felt edits improved the accuracy of the model.

The model consists of 3 convolutional layers and 3 maxpooling layers. Two dropout layers were used where 30% of the neurons were randomly shut down.

The input matrix of our model was a matrix of dimension (7356*162). Here each row represented one data point and the 162 columns in each row are simply the features extracted stacked horizontally.

The results

The parameters such as the dropout percentage, training length in terms of the number of epochs, batch size and learning rate was adjusted to get the highest possible accuracy at our validation set.

The following parameters were set:

Batch size=256, epochs=90, learning rate = 0.0000001.

We achieved the best result at the 68th epoch where out model had a training accuracy of 78.41%. The losses at this epoch was 0.7978.

The following figures were generated that further takes our cause:

Figure 1: Losses versus Epochs trained

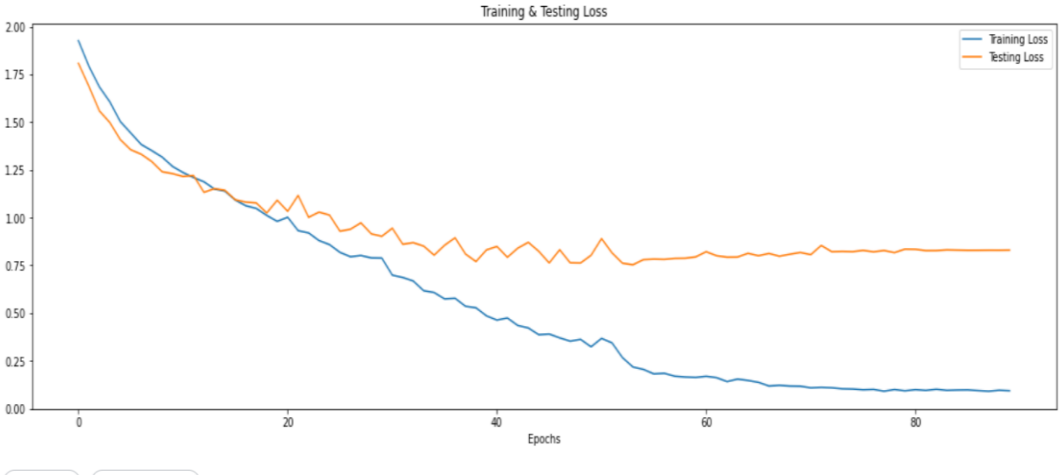

Figure 2: Accuracies versus Epochs trained

As we can see our accuracy peaks at the 68th epoch where is has a value of 78.4% (testing accuracy).

The accuracy on the validation set is pretty good considering accuracy of the existing models on the RAVDESS dataset.

Brief comparison with other models

Table 2: The Widely used pre existing models performance on Ravdess dataset

Model	Accuracy
SVM	0.791
Random Forest	0.634
Gradient Boosting	0.616
KNN	0.443
Decision Tree	0.342
VGG16	0.747
VGG19	0.763

*This info table has been taken from a paper Speech emotion recognition in neurological disorders using Convolutional Neural Network Sharif Noor Zisad, Mohammad Shahadat Hossain1, Karl Andersson.

Among the existing models that predict on the RAVDESS dataset for audio only files, SVM has the highest accuracy of 79.1% and the second one in terms of accuracy is VGG19 with the accuracy of 76.3%. Our model in this project had an accuracy of 78.4% on testing which is a comfortable second among the widely used models of today. It is also worthy to mention that our validation set also contained some augmented speech samples, and hence a 78.4% accuracy must is a lower (worst case) estimate of this models capability.

The basic architecture of the model that we used in our project was said previously to be taken from the Kaggle expert, Shivam Burnwal. He attained an accuracy of 61% with the model which he deemed to be quite good for the RAVDESS dataset.

Our model was produced via some small changes in his architecture. However, by making vast changes in hyper parameter tuning and bringing out major changes in audio feature extraction we pushed the model to reach its' full potential. In the process, we improved the accuracy from approximately 61% to approximately 78%.

Conclusion:

We have seen that our new model is performing as good as the top existing model on the audio only speech emotion recognition in the RAVDESS dataset. Briefly, we took a Kaggle expert's model that had average performance and improved it via mainly hyper parameter tuning to increase its' accuracy by approximately 17%. This model now, is performing as good as the well established models in the field.

Due to the lack of time and the ever approaching deadline we are not able to tune the hyper parameters further to push its' accuracy up by a couple of percentages which will make the model best among the pre existing and well established models. We are confident that our architecture is yet to reach its' full potential and although we are submitting it now as our project, we shall continue to work on it to make sure it does

References:

Database:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) : https://zenodo.org/record/1188976

Paper mentioned:

Speech emotion recognition in neurological disorders using Convolutional Neural Network Sharif Noor Zisad, Mohammad Shahadat Hossain1, Karl Andersson.

User mentioned:

Shivam Burnwal | Expert | Kaggle.