Identification of Foreign Language from Speech

Mr. Fakruddin Mohammed
fam281@g.harvard.edu

Problem

There were number of studies on language identification from the speech files using the CNN deep learning models. These studies have used disparate languages in terms of geography/regional (Asian, European, UK etc), dialects and accents. However, none of the studies have focused in establishing the effectiveness of the deep learning models when almost similar sounding (dialect and accent) languages are fed to the model. In this study, I am perusing the effectiveness of CNN model in detecting the languages spoken: Arabic, Arabic Egyptian and Arabic Sudanese using the Tensorflow & Keras technology. The data set downloaded from the TopCoder website is used for this study.

Data Set

The data set was downloaded from the TopCoder website (`https://community.topcoder.com/longcontest`). The data set comprises of 10 seconds audio recordings in 176 languages. The size of the data set is 4.6GB with 66,176 files for training and 12,320 files for training.

Models & Techniques

CNN Three Convolution layer neural network model was used for this study. The CNN model is run on the speech files converted to MFCC and Melspectogram.

Melspectogram The effectiveness of CNN model is tested by converting audio speech files to Melspectograms images of size 32x32 and 64x64.

Mel Frequency Cepstral Coefficients (MFCC) The audio speech files are converted to MFCC vectors of sizes 11, 24 and 32 and effectiveness of CNN model is examined.

Technology

Hardware The model was run on local computer with 8GB RAM. The model execution time is between 5-50 minutes depending on the size of the input feature vectors.

Software

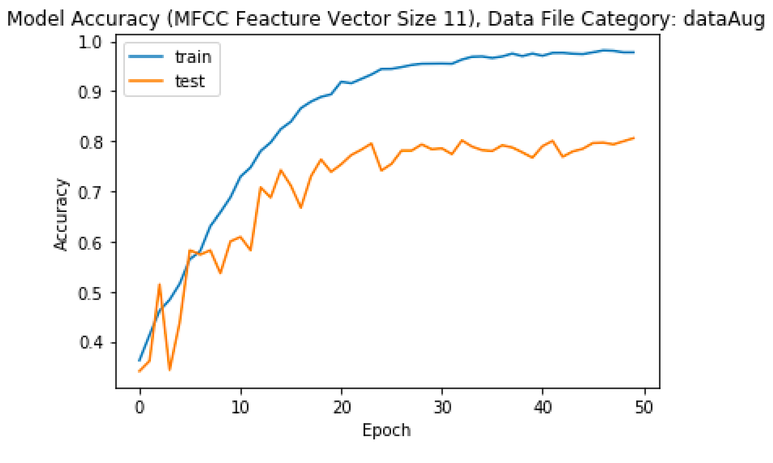
- Python 3.6
- Tensorflow & Keras
- Librosa, Numpy, scipy, shutil, io, matplotlib etc.

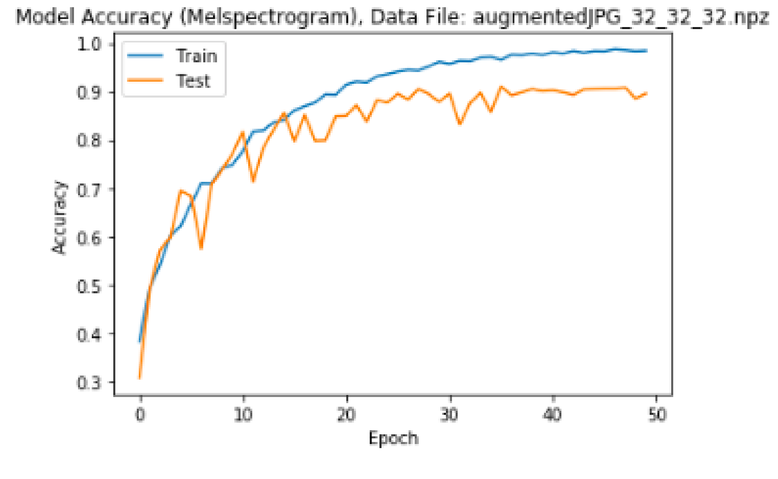
Experiments

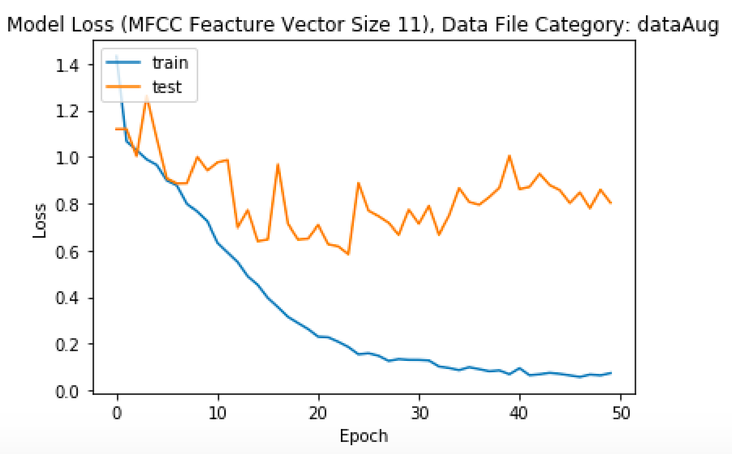
Dataset	Train/Test Split	Technique	Accuracy
Original Data	902/226	MFCC(11)	0.6
Original Data	902/226	MFCC(24)	0.6
Original Data	902/226	MFCC(32)	0.6
Original Data	902/226	Melspectogram(32x32)	0.7
Original Data	902/226	Melspectogram(64x64)	0.8
Original Data with Augmentation	4512/1128	MFCC(11)	0.8
Original Data with Augmentation	4512/1128	MFCC(24)	0.8
Original Data with Augmentation	4512/1128	MFCC(32)	0.8
Original Data with Augmentation	4512/1128	Melspectogram(32x32)	**0.9**
Original Data with Augmentation	4512/1128	Melspectogram(64x64)	**0.9**

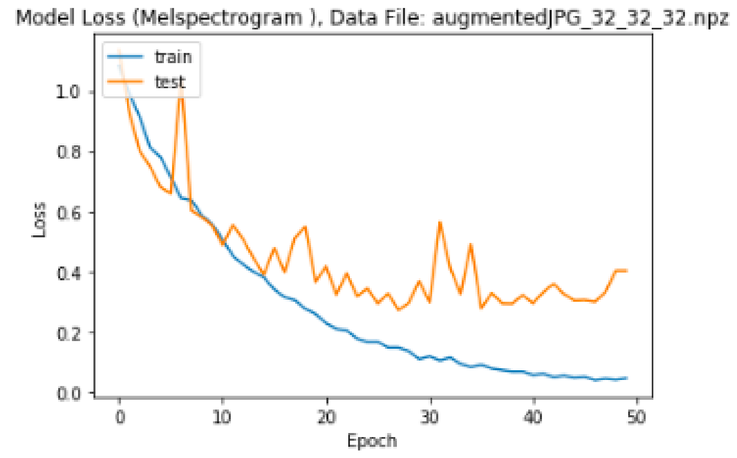
Results, Conclusions and Next Steps

Results









Conclusions

- The results shows, the CNN model performs reasonably well giving accuracy of 90% even when similar sounding, dialect languages are fed to the model. However, this accuracy is slightly less than reported accuracy in literature.
- The results shows audio speech files converted to spectrogram is giving superior performance compared to MFCC.
- Increasing the size of the spectrogram or MFCC vector sizes had a very little impact on model accuracy.
- The results also shows that, relying on data augmentation is an extremely useful tool when the input data set is too low.

Next Steps

- Even though the model accuracy is around 90%, I wanted to explore if there are any other CNN architectures that give better accuracy. For example explore more deeper architecture like VGGNet, ResNet etc.
- Test the model prediction stability using the k-fold cross validation techniques.
- Experiment with the RNN/LSTM model, as the identification of language is dependent on context and the RNN models are best suited for it.
- The literature survey reveals that, in addition to MFCC, using their 1st and 2nd order derivatives as feature vectors improves the model performance

YouTube URLs

- 2 minutes video: `https://youtu.be/8jNt1EIKpxc`
- 15 minutes video: `https://youtu.be/FFq7wMZ5fnY`