Deep Learning for Spech-Rap-Singing Audio Classification

Axel Ind, Msc. Computer Science

July 14, 2021

Abstract

I have implemented two different deep learning architecture to learn the 3-output classification of 3 second music clips with a $16kHz$ sample rate. Implementations make use of Librosa for audio-preprocessing, and the Keras environment for model implementation. Labels are learned with a 13-variable MFCC as input data. Results achieved are: 98%, 89% accuracy on test data for the CNN and MLP respectively.

1 Audio-Preprocessing

1. Confirmed that each sample match the three second length and $16kHz$ sample rate.
2. Used `Librosa` to extract signal data from each sample.
3. Used `Librosa` to convert signal data to 13-variable MFCC.
4. Stored data in a `.json` file with format: ('label':{...},'MFCC':{[...]}).

2 Architectures

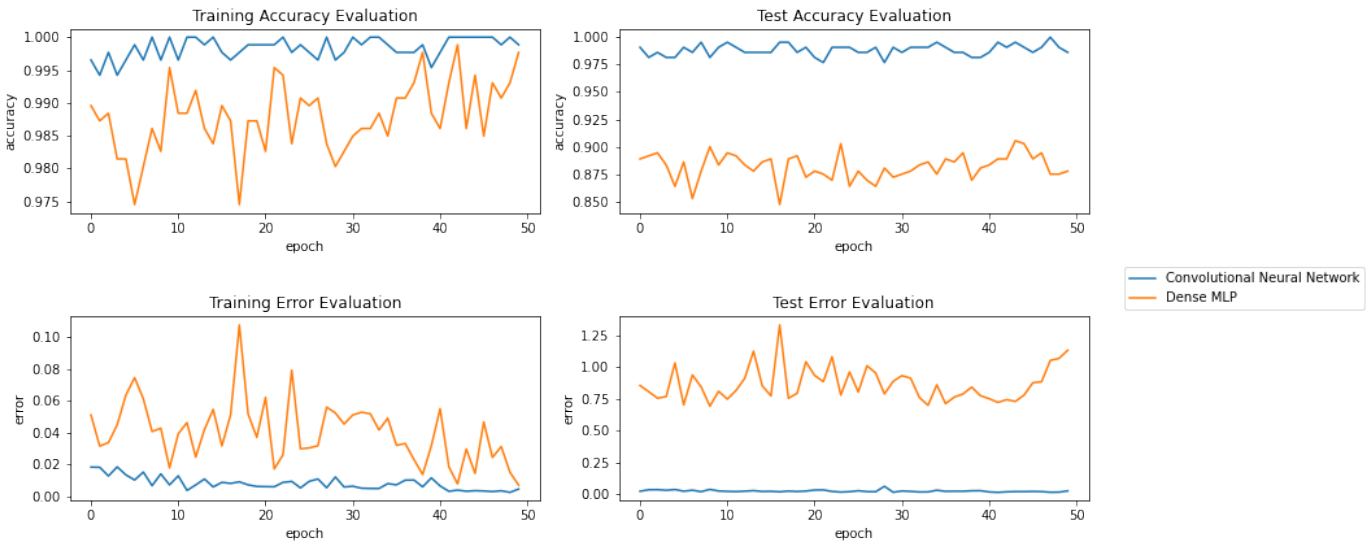
2.1 Classic Dense Network

- Layer 1: Flatten Layer (to reduce MFCC to vector)
- Layer 2-4: Dense Layers (relu, dropout) with 512, 256, and 64 nodes, respectively.
- Layer 5: Output Layer (3 nodes, softmax)

2.2 Convolutional Neural Network

- Layers 1-6 alternate this standard pair of convolution-oriented layers giving a total of three Convolutional Layers:
  - a: 2DConvolution (32 filters, 3×3 kernel size, relu)
  - b: MaxPool (pool_size=3×3, strides=2×2, relu) with Batch Normalisation.
- Layer 3: Dense Layer (relu, dropout) with 64 nodes.
- Layer 4: Output Layer (3 nodes, softmax)

3 Results



3.1 CNN Results

With $\mu = 0.0001, \bar{t} = 5.42s$

- **Average Accuracy**: 94.657%
- **Max Accuracy**: 99%

3.2 Dense Network Results

With $\mu = 0.0001$, $dropout = 0.1, \bar{t} = 4.58s$

- **Average Accuracy**: 89.657%
- **Max Accuracy**: 95.735%

4 Limitations

- Training can currently include different features from the same person's voice as the test set.
- The amount of training data currently available is rather limited. Perturbating the existing data could potentially increase accuracy.
- Hyper-parameter optimisation will be necessary.
- An LSTM solution was implemented but does not run in `Python 3.8` and I am still troubleshooting compatibility.