Deep Learning for Spech-Rap-Singing Audio Classification

Axel Ind, Msc. Computer Science

July 15, 2021

Abstract

I have implemented three 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: 92%, 81%, and 70% accuracy on test data for the CNN, MLP and LSTM 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)

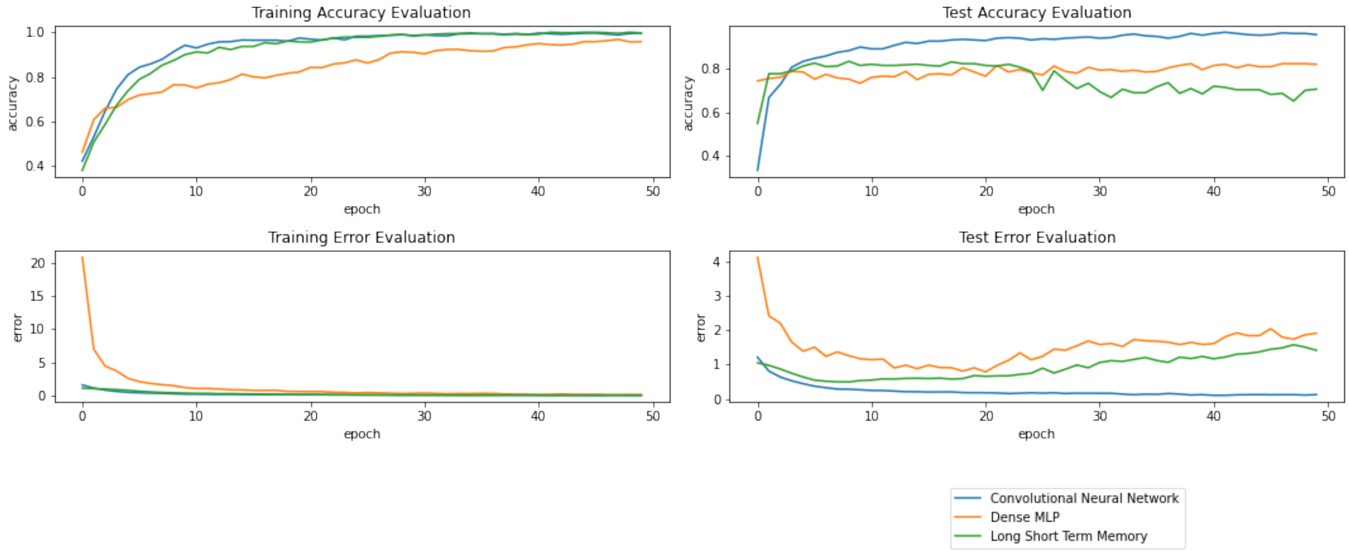
2.2 Convolutional Neural Network

- 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)

2.3 Long Short Term Memory Network

- Layer 1-2: LSTM (tanh) with 64 nodes.
- layer 3: Dense Layer (relu,dropout) with 64 nodes.
- Layer 4: Output Layer (3 nodes, softmax)

3 Results



3.1 CNN Results

With $\mu = 0.0001, \bar{t} = 3.23s$

- **Average Accuracy (preserved samples**: 92.31%
- **Average Accuracy (randomised samples**: 94.56%

3.2 Dense Network Results

With $\mu = 0.0001, dropout = 0.1, \bar{t} = 13.15s$

- **Average Accuracy (preserved samples**: 81.72%
- **Average Accuracy (randomised samples**: 86.657%

3.3 LSTM Network Results

With $\mu = 0.0001, \bar{t} = 48.26s$

- **Average Accuracy (preserved samples**: 70.43%
- **Average Accuracy (randomised samples**: 92.31%