Genre Detection with Deep Neural Networks

Matt Jones *mattjone@stanford.edu*
Daniel Way *dway91@stanford.edu*
Yasaman Shirian *yshirian@stanford.edu*

CS 230 Winter 2019

Problem

Motivation:

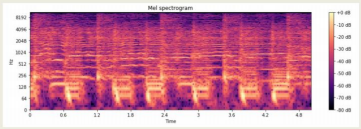
New songs demonstrate how genres can creatively borrow from each other to create invigorating new, yet familiar styles of music. While much previous work has been put into music genre classification, very little have explored this emerging phenomenon more deeply.

Objective:

We took two parallel approach to tackle music tagging task with NN:

- To achieve high performance on music genre-detection with proposing a fine tuned model after explorations with different architecture and optimization method choice.
- To propose a novel architecture consisting of CNN and fully connected layers with aggressive search of optimal choice of hyper-parameters(learning rate, clip length,..)

Dataset



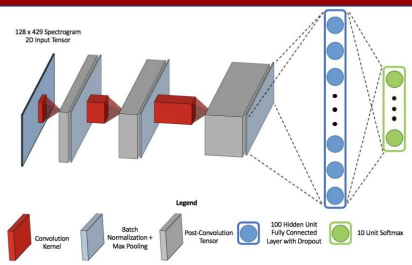
The figure above represents a typical mel-spectographic representation of the audio.

Horizontal axis is time and vertical axis is frequency, with the intensity of the color denoting loudness at any particular instance in the data.

Each audio-clips are 30 seconds long. Audio-clips are pre-processed into 1366 frames with sampling frequency of 256 consisting of 96 frequency mel bins.

Dataset: GTZAN, FMA

Novel Architecture



Our network consisted of a convolutional and a fully connected section.

The CNN section has 3 layers. Each convolutional layer consists of a 5x5 kernel with 16, 32, and 64 filters respectively followed by a batch normalization and max pooling layer.

The fully connected section has a 100 hidden-unit layer with tanh activation and 50% dropout and a final 10-unit layer with softmax for final classification.

Results - Novel architecture

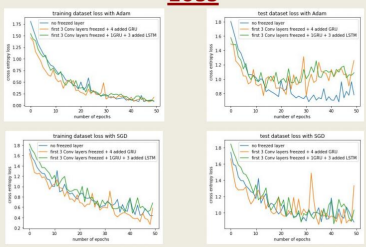
Learning Rate	Normalization	Training Accuracy	Validation Accuracy
.0005	No	20%	20%
.0001	No	99.5%	43.3%
.00005	No	99.5%	33%
.0005	Yes	13.7%	13.3%
.0001	Yes	99.5%	33%
.00005	Yes	99.5%	40%

Clip Length	Train / Val Size	Paremeters	Training Acc.	Validation Acc.
30 sec	500 / 50	5,104,714	99%	38%
30 sec	900 / 50	5,104,714	99%	46%
10 sec	2700 / 150	1,664,714	96%	76%

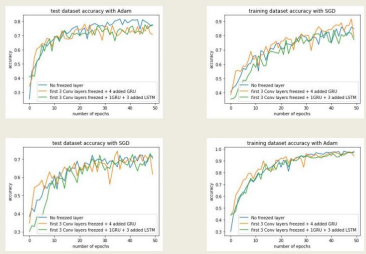
Reducing the audio segment size had the most substantial impact as it both increased the number of samples and significantly decreased the number of parameters in the model

Results - fine tuning

Loss



Accuracy



Original model is proposed by [1] and consists of 4 stacked [Conv layer , BN, relu,dropout] and followed by 2 GRU.

We could achieve accuracy of 85% for successfully tagging 7 music genres, our fine tuned model could outperform the pre-trained model by [1].

Adam optimization method has better performance in minimizing the loss.

Convolutional layers at the beginning of the NN have more impact on achieving higher test accuracy than RNN layers (GRU and LSTM) added at the depth of the NN.

References

[1] A. Jiménez and F. José. Music genre recognition with deep neural networks.

[2] S. Oramas et al. Multimodal deep learning for music genre

[3] K. Choi, G. Fazekas, and M. Sandler. Automatic tagging using deep convolutional neural networks.

[4] H. Cohen and C. Lefebvre. Handbook of categorization in cognitive science.