Dilated CNN + Classical Algorithms for Music Genre Classification

Haojun Li[1], Siqi Xue[2], Jialun Zhang[2]

[1] Department of CS, Stanford University, [2] ICME, Stanford University

Abstract

Our goal is to improve **classical algorithm**'s performance in classifying music genres with a **dilated convolutional neural network**. First, we established baselines with classical algorithms with minimally pre-processed audio data and evaluate their performance. Then, we trained a dilated CNN, and use the different layers of our pre-trained CNN as the feature input for a few of the classical algorithms. The CNN has improved training time for other algorithms, but has proven to be easily overfitting the data. Classical algorithms seems to have some regularizing effect, but only to some extent. In general they achieve better results when working together.

Data and Preprocessing

Our data set consists of 10 genres of music files. We will be only using 5 genres that we selected, namely classical, hiphop, metal, pop, blues. Each genre contains 100 pieces of music in wave form. Using the LibROSA library in Python, the data is preprocessed into the MFCC features, which allows us to represent each file as a 2D Numpy array.

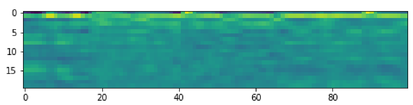

Figure 1: Classical MFCC Features

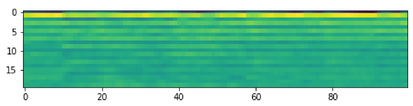

Figure 2: Metal MFCC Features

Dilated CNN

The convolution neural network that we trained consists of 2 dilated convolution units. Exact architecure is shown below:

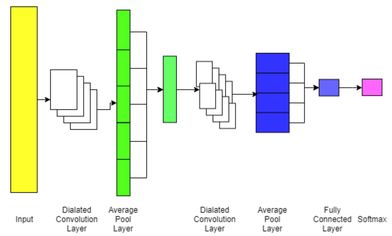

Figure 3: Dilated CNN Architecture

The Dilated CNN we have trained is very small and by itself it achieved around 86% train and test accuracy.

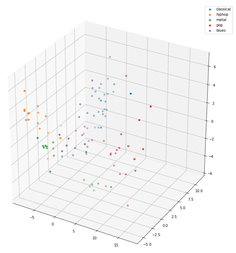

Figure 4: DCNN Layer 1

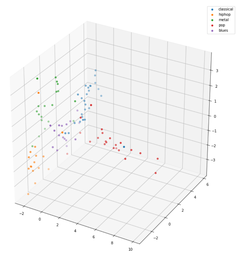

Figure 5: DCNN Layer 2

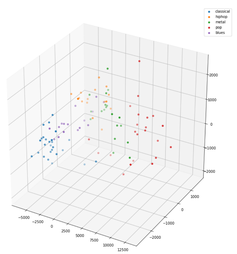

Figure 6: Raw PCA

Graph above shows PCA of second layer features have significantly separated out the pop songs comparing to Layer 1, and even better than PCA analysis on Raw data.

Classical Algorithms

We investigate the performance of four classical algorithms: Logistic Regression (LR), GDA, Random Forests and SVM. For each algorithm, we use three sets of inputs: (1) the input of the Dilated CNN network (2) the first dilated convolution layer in 4 (3) the second dilated convolution layer in 5. A comparison of the LR and CNN results using Layer 1 is shown in 7 and 8

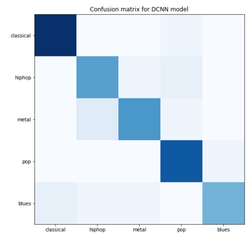

Figure 7: CNN ConfusionMatrix

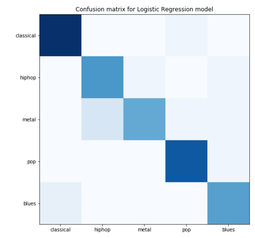

Figure 8: LR confusion Matrix

As we can see in the above figures, there are only subtle differences in the coloring of the labels. Interestingly, Logistic regression actually performed better than the CNN with an accuracy of 88%, so we suspect some regularization effect by combining CNN and classical algorithms. However (below in PCA analysis), some purples are embedded in blue labels, which means classical algorithms are still constrained by CNN's overfitted features.

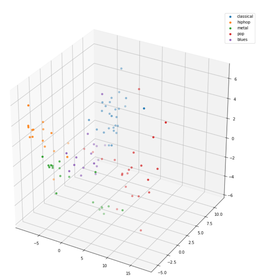

Figure 9: True Labels PCA

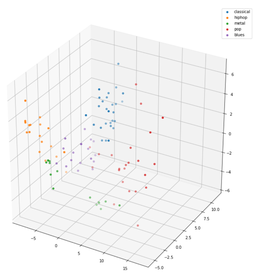

Figure 10: Logistic Regression PCA

Table 1: Accuracy Comparison

	LR	GDA	Random Forest	SVM
raw input train	1.0	0.8875	1.0	1.0
raw input test	0.77	0.77	0.76	0.28
layer 1 train	0.9875	1.0	0.99	0.92
layer 1 test	0.88	0.61	0.86	0.87
layer 2 train	0.945	0.94	0.98	0.94
layer 2 test	0.88	0.8	0.87	0.84

Discussion

We see that when training with raw data, the classical algorithms not only takes much longer, but also over-fit the data by a lot. Both aspects are greatly improved by incorporating CNN features. We suspect that combining classical algorithm and CNN have some regularization effect as LR with features from the CNN has greater accuracy. However, we also realize that classical algorithms are constrained by how well the CNN extract features, as shown in PCA analysis.

Future Work

Convolutional neural networks might not be the best architecture for music classification. RNN architectures such as GRU and LSTM would likely produce better accuracy, but futher work needed to make them feature extractors. With more data we can also train deeper networks and also reduce overfitting.

Reference

[1] G. Tzanetakis and P. Cook, GTZAN Genre Collection `http://marsyasweb.appspot.com/download/data_sets`

[2] Van Den Oord, Aäron, et al. "WaveNet: A generative model for raw audio." SSW. 2016.