Music Genre Classification Using Mel Spectrogram Representations

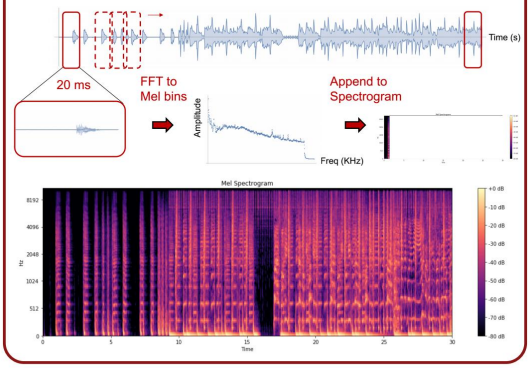
Aidan O'Beirne (aidano@stanford.edu), Alvaro Zamora (alvarozamora@stanford.edu)
Stanford University

Overview

The objective of this project is to create and train a deep neural network to classify music into the following categories: Classical, Jazz, Metal, Pop, Country, Blues, Disco, Metal, Rock, Reggae and Hip-Hop. The inputs to the neural net are two dimensional Mel spectrograms of 3 second long time domain audio signals. We used a CNN architecture composed of 4 convolutional layers, 1 fully connected layer, and 1 softmax layer containing 10 neurons for each of the 10 different genre classes. We trained the CNN using Adam optimization to minimize a categorical cross entropy loss function.

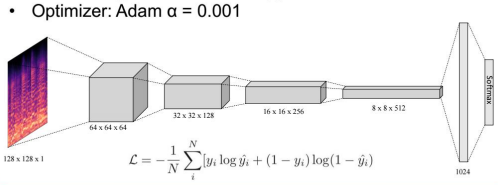
Data and Preprocessing

We collected 1,000 (100 of each genre) 30 second music files from the GZTAN open repository and preprocessed them into 10,000 3 second long audio clips, and then converted these clips into Mel spectrograms. We divided the data into the following sets: 95% train, 2.5% dev, and 2.5% test.



CNN Model Architecture

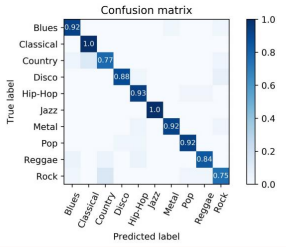
- Square inputs of 128 frequencies x 128 time steps
- Each convolution is a 2 x 2 filter with a stride of 2:1:2:1 for layers 1:2:3:4
- Each maxpool is a 2 x 2 filter with a stride of 1
- Initialization: Xavier
- Activation: Relu, except for the softmax layer
- Regularization: L2 λ= 0.019, BatchNorm
- Loss: Categorical Cross Entropy
- Optimizer: Adam α = 0.001



$$\mathcal{L} = -\frac{1}{N}\sum_{i}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)$$

Results

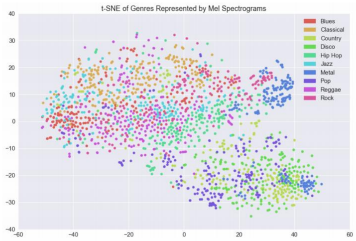
The best trained net performs at **90% test set accuracy** and **91% dev set accuracy.** The best trained net exhibits variance, as the train set accuracy is 99%. This discrepancy is likely due to having a small dataset as standard regularization techniques only reduced performance.



Model	Accuracy
Elu Activation	83%
Dropout (p = 0.9)	78%
L2 Regularization λ = 0.002	86%
Maxnorm (max = 3)	89%
Best (λ = 0.0019)	91%

Discussion

The model had the most difficulty classifying Country and Rock (77% and 75% accuracy, respectively). It classified Jazz and Classical easily (100% accuracy). From the t-SNE diagram, we see that Bayes' error for classifying each genre may be different.



Future Work

With more computational resources, we can expand dataset to include over 100,000 songs from the FMA dataset as well as attempt using larger CNN architectures. We may also try using other inputs along with Mel-spectrograms, such as linear predictive ceptra or MFCC's. Music genre style transfer using standard neural style transfer and Griffin Lim algorithms for spectrogram inversion is also a long-term goal.

References

1. Tao Feng. Deep learning for music genre classification
2. Lee, Chang-Hsing, et al. "Automatic music genre classification using modulation spectral contrast feature." Multimedia and Expo, 2007 IEEE International Conference on. IEEE, 2007.
3. Laurier, Cyril, et al. "Audio music mood classification using support vector machine." MIREX task on Audio Mood Classification (2007)
4. Karpov, Igor, and Devika Subramanian. "Hidden Markov classification for musical genres." Course Project (2002).
5. Despois, Julien. 2016. Finding the genre of a song with Deep Learning
6. Perraudin, Nathanaël, et al. (2013). A Fast Griffin-Lim Algorithm. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics .
7. Griffin, Daniel and Lim, Jae. (1984). Signal Estimation from Modified Short-Time Fourier Transform
8. Ahrendt, Peter. 2015. Decision time horizon for music genre classification using short time features.