# Classification of Neural Network Generated Audio Files

Laurel Doyle (ldoyle@uw.edu) and Rebecca Alexandra Lopez (rlopez01@uw.edu)
Department of Applied Mathematics, University of Washington

## Motivation

Much exploration has been done into the application of neural networks to image tasks, but quite a bit less has been done in the field of understanding audio. In our project, we aim to utilize the GTZAN music data set for generating audio files of specified genres. We then aim to classify these generated files appropriately with a neural network trained on the original data set. We implement a recurrent neural network for the generation task as well as a convolutional network for the classification task. Our goal was to synthesize the material learned in the Applied Mathematics course on Inferring the Structures of Complex Systems and to explore various forms of NN implementation and application.

## Background

Understanding audio data continues to present a challenge in the field of deep learning due its high dimensionality. Often audio is converted to images through spectrograms or wave plots to utilize the great success in established architectures for classifying and generating visual information.

For our deep learning project, we develop two networks to explore music generation and classification tasks. The first network would generate music of a specified genre using recursive techniques on the sequential audio training data. The second network would utilize image classification mechanisms and convolutional network architecture to process the generated music data. This network would be trained on the original dataset used to train the audio generation network.

For these tasks, we chose to use the GTZAN Dataset which features a variety of audio from 2000-2001. This audio is divided into 10 genres with 100 thirty second music clips each.

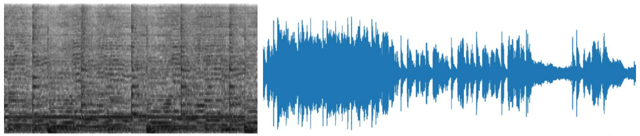


Figure 1: This figure displays two potential visual presentations of the same audio file: (Left) Spectrogram of Rock music clip, (Right) Wave plot of Rock music clip.

We utilize various network types and layers :

- Fully connected: Feeds an input forward by multiplying it by a weight vector and adding a bias
- Convolutional: Apply filters to an input in order to create a feature map that summarizes presence of detected features in such input
- Recurrent: Utilizes past memory to feed forward an input
  - Long Short Term Memory: Learn long-term order dependence for sequence predication problems using input, forget, and output gate
  - Gated Recurrent Unit: Uses previous hidden states on input to produce output using reset and update gate
  - Bi-directional: Connect two hidden layers of opposite directions back to the same output

## Generation

The network for generating music is a combination of convolutional and recurrent neural networks. The music files were sampled at a rate of 10,000 Hz and 90 of each genre were used for training with 10 left for testing. Training batches consisted of 29 one-second long clips, each from the same song. (The last second of the thirty seconds is used for validation during training.) 9/10 of a second of music is used to predict the following 1/10 second of music, as shown in Figure 2.

A 1D convolution is applied to the training song clip and then the resulting vector is processed in a GRU recurrent layer with bidirectional information flow. This information is outputted in a vector of length 1,000 after passing through a linear layer. This is the prediction for the following 1/10 second of music.

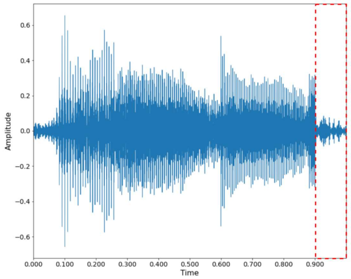


Figure 2: Audio file from the training data set with the seed audio of 0.9 seconds length and the generated output of 0.1 seconds length highlighted in red.

Audio files of 5 seconds in length are then generated using 9/10 second of seed audio from the testing set. The network is applied iteratively to produce the next 1/10 second of audio. These outputted files are shown in Figure 3, and represent the information we hope to classify using our classification network.

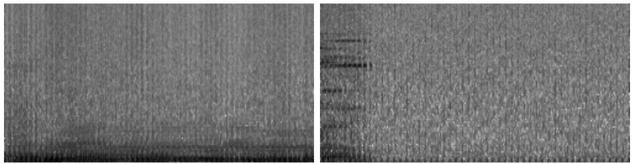
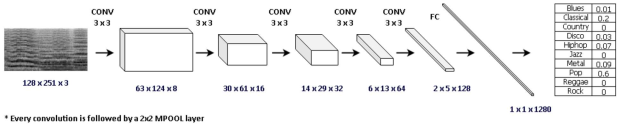


Figure 3: Mel spectrograms for two generated audio files: (Left) Spectogram for generated 'pop' song, (Right) Spectogram for generated 'jazz' song.

## Classification

In order to now classify the generated music clips, we convert our audio files into visual files.
Once converted to mel spectograms, we utilize a neural network with five convolutional layers and a fully connected layer.
In terms of parameters, we found the Adam optimizer to work best with a learning rate of 0.0001. Additionally, we introduce batch normalization to increase the stability and speed of our network.



* Every convolution is followed by a 2x2 MPOOL layer

Figure 4: Network architecture for the music genre classification task.

## Results

Below, we see our validation accuracy as a function of epochs on the validation data.

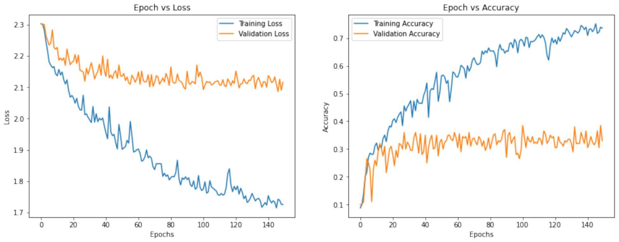


Figure 5: Diagnostic plots for our classification model : (Left) Validation accuracy as a function of epoch count, (Right) Training loss as a function of epoch count.

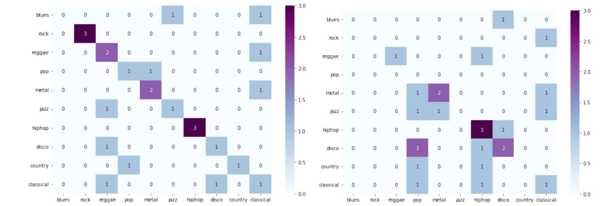


Figure 6: This figure displays the confusion matrices of our predicted versus correct labels: (Left) Model done on all genres for training set, (Right) Model on generated audio.

## Selected Bibliography

1. "Music-Generation: Music Generation with Variational Autoencoders (VAES), with a Input Data of the Music from Disco Genre." GitHub, https://github.com/cjw531/ee435-music-generation.
2. Goodfellow, Ian, et al. Deep Learning. MIT Press, 2017.
3. Lewinson, Eryk. "Implementing Yann LeCun's Lenet-5 in Pytorch." Medium, Towards Data Science, 9 May 2020, https://towardsdatascience.com/implementing-yann-lecuns-lenet-5-in-pytorch-5e05a0911320.
4.Music Classification and Generation with Spectrograms - Neuromatch. https://deeplearning.neuromatch.io/projects/ComputerVision/spectrogram_analysis.html.
5. Tham, Isaac. "Generating Music Using Deep Learning." Medium, Towards Data Science, 30 Aug. 2021, https://towardsdatascience.com/generating-music-using-deep-learning-cb5843a9d55e.

We would like to acknowledge and thank Dr Eli Shlizerman, Ryan Vogt, & the Applied Mathematics Department at the University of Washington.