Landon Buell

Jan 2023

Audio Classification of Musical Instruments with a Multimodal Neural Network

1. Abstract
2. Introduction

Signal classification is a broad area of signal processing that seeks to group waveform-like objects into categories based on properties within the signal. From musical

Consider a library of synthetically generated signals that represent digital audio waveforms that we would like a associate with a real-world musical instruments. If there were only a few dozen or hundreds of samples, it’s feasible to have a human manually open each file as an audio track, listen to it, and then determine what instrument it most sounds like. However, this process is intensive and very subjective when it comes to producing a label. Listeners that grew up in different regions of the world, exposed to different musical styles, played across multiple instruments may produce wildly inconsistent conclusions when analyzing the same audio sample. This, combined with a library of waveforms that may number in the thousands or millions makes it impossible for even a group of humans to complete in a reasonable amount of time with reasonable accuracy of category assignment.

Because of the constraints that humans add, it makes sense to attempt to automate this process with the aid of a computer, but audio classification from a waveform is not a task that classical computer algorithms are known to excel at [citation needed]. To succeed, we need to combine the computational efficiency, with the intelligence of a human brain. For this, we look no further than a *neural network.*

A neural network is a mathematic function that accepts a collection of numerical inputs that encode a particular sample to produce a set of outputs that encode a decision. In the cause of digital audio classification.

1. Feature Selections

Neural networks are mathematical functions that use a collection of inputs called *features* to produce outputs called *predictions* [citation needed]. Features are compact-low dimensional representations of a given input which must effectively capture the key characteristics of their parent sample. Features with a category or class should behave similarly and features between classes should behave differently. For example, if characterizing cats against dogs, we might use weight since dogs in general are heavier than cats, whereas using the number of legs would not be very useful as cats and dogs typically have four legs.

1. Spectrogram Features

A spectrogram is a 2D representation of a waveform that characterizes a signals spectral evolution over time. A spectrogram is built by first dividing up a waveform into a collection of overlapping slices of time called *analysis frames* and then applying a Discrete Fourier Transform to each frame, and then aggregating the frames back together. The result is a 2D matrix S with entry *S*[*i,j]* encoding the energy of the waveform at time frame *i* and frequency bin *j*, which also allows us to represent a 2D image as opposed to a 1D signal.

1. Time-Domain Features

Features drawn from the time-domain parse the waveform in its amplitude vs. time representation. Note that for this set of features, the actual waveform itself is never presented to the neural network, it is only used to build up the collection of features with are then presented to the neural network.

1. Time Domain Envelope
2. Zero Crossing Rate
3. Temporal Center of Mass
4. Auto Correlation Coefficients
5. Frequency-Domain Features

Features drawn from the frequency domain parse the waveform in its energy vs. frequency representation. Note that for this set features, the actual frequency spectrum

1. Mel Frequency Cepstrum Coefficients
2. Frequency Center of Mass
3. Multimodal Architecture

We have chosen a collection of features derived from both the time and frequency domain of a digital signal. The value for each feature in this format represents a single, independent property of the sample. Moreover, each feature can stand by itself because it does not directly rely on any other value in the collection. For this reason, we can aggregate each feature from the time and frequency domains and concatenate them into a single *feature vector.* This is a (1 x p) array that neatly summarizes the characteristics of the sample in a sort of list-of-describing-qualities format.

We have also chosen a spectrogram to represent each sample in an image-like format. image. In the spectrogram matrix, any entry S[i,j] represents the energy of frequency bin “I” at a particular time frame “j”, but taken out of context, is not a particularly useful piece of information.

Unlike the previous feature vector, the elements of the spectrogram matrix are not as independent and carry the most meaning when a part of the larger whole. Furthermore, the arrangement of the elements in this matrix are crucial to representing the sample due to the nature of what the spectrogram physically encodes. For example, S[I,j+1] relates to S[I,j] in that the former shows how the energy related to the latter one time frame later.

The problem then arises that these are to different *modalities* of the same sample. Both the (1 x p) feature vector and then (m x n) spectrogram encode information derived from the same source, but in incompatible formats. This can be though of like a driver’s license where one side will contain a list of qualities such as height, weight, hair color, and eye color, and is paired with an image that was taken on the day it was issued. Both ideas describe the driver, but in two different ways. To take full advantage of both modalities, we must build a *multimodal neural network* which can accept both formats, learn a set of weights that allow for classification, and then use the combined inputs to generate a single prediction label.

1. Performance Evaluation Techniques
2. Performance of models