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 digital 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]

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 most cats and dogs have four legs.

1. Multimodal Architecture
2. Performance Evaluation Techniques
3. Performance of models