Landon Buell

14 December 2020

Abstract

Classifying audio signals with machine learning has become an important topic of research in the past few years. Models often involve the input of a 2-D spectrogram or 1-D feature vector into a unimodal network such as a Convolutional Neural Network (CNN) or Multilayer Perceptron (MLP). In this study, we explore automatic classification of musical instruments using new hybrid neural-network architecture that combines the CNN and MLP models and provides superior performance over models that rely solely on one or the other. This hybrid network uses two branches, one being a CNN to process an image-like 2-D spectrogram, and the other being an MLP to process a 1-D feature vector. Within the model, a hidden layer combines activations from the two branches by concatenating them into a single 1-D dense layer, thus any predictions are a product of both branches. We describe in detail the creating of the spectrogram and features, as well as how they influence the chosen network architecture. We finish with a practical demonstration that uses this classifier model to match waveforms from a chaotic music synthesizer to real-world musical instruments. Training data is from studio recordings of the Philharmonia Symphony Orchestra and University of Iowa's Electronic Music Studios

Musical Instrument Classification Using a Hybrid Neural Network

Abstract, See above - Completed, may need editing

1. Introduction
   1. Introduce & Describe the task
      1. Map audio files to source musical instrument
      2. Humans are good at matching sounds to sources
      3. Can a computer do it?
      4. Difficult for most traditional programming techniques
      5. Classification task
   2. Outline parts of the task
      1. We use a neural network to do this
      2. Neural Network requires inputs called features
      3. We use physics, DSP to derive features
      4. Two modes of information: Spectrogram & feature vector
      5. We design our network to account for the two modes
2. The Neural network
   1. Introduction & Structure
      1. Math Function: Inputs -> Outputs
      2. Input is transformed by functions called layers
      3. Directed Computational graph controls flow of information
   2. Input and Output and Training
      1. Stimulus input -> Source prediction
      2. Array input -> Source prediction
      3. Inputs are features – this is what we need to develop
      4. Outputs are predictions
3. Features and The Multimodal Model
   1. Describe features as needed
      1. Low Variability / High variability
      2. Compatibility of 1D + 2D inputs
   2. We develop the following features
      1. 24 features from time & freq. space
      2. Spectrogram matrix
      3. Detail physical nature of features
      4. Spectrogram shows success with CNN
      5. Feature vector shows success with MLP
      6. We can combine the two models to generate an aggregated prediction
   3. Use two *modalities* of features
      1. spectrogram – 2D array
      2. Feature vector – 1D array
      3. Represent the same waveform in a different way
      4. Non-compatible – we can’t just *smush* them
   4. Multiview learning
      1. Multimodal/multirepresentation learning has been explored
      2. We find improved performance
      3. Compare results from unimodal models to hybrid network
4. Experimental Classification Results
   1. Using K-Folds X-Validation
      1. Compare confusion matrices
      2. Compare metric scores
      3. What did we learn from X-Val? Consistency, generalization?
   2. Report of Metrics
      1. Loss, Accuracy, Precision, Recall scores (Figs already made)
      2. Metrics over a period of training
      3. Metrics on Validation set
   3. Final Classifications
      1. Compare Features/ Spectrogram
      2. Compare a few unlabeled examples
5. Conclusion?
   1. The Classifier Works!
      1. We show high classification performance w/ confusions
      2. Compare classification metrics across models
   2. Multimodal, on average does better than unimodal
      1. Compare modes (spectrogram does pretty well in some cases alone?)
      2. Why does multimodal work better?

Additional notes about this outline:

Romain Numerals (I. II. Etc.) represent sections

Uppercase Latin letters represent subsections

Arabic numerals represent paragraphs or groups of paragraphs

Lowercase Latin letters represent sentences or groups of sentence