COMP150 DR Final Project Proposal

Ari Brown and Julie Jiang

October 26, 2017

Abstract

The problem of sound classification has been studied in depth and has multiple applications related to identity discrimination, enhanced hearing aids, robotics, and music technology. There are two ways in which the problem of sound classification can be approached. The first method is empirical in nature, and requires a database of sounds to learn from through feature extraction. The second method is more top down, by using predefined feature rules to make classification decisions. In this project, we take the developmental approach, and use learning mechanisms to enable a robot to gain meaningful information about its environment through sound. Specifically, we compare two classification algorithms: a k-nearest neighbor characterization of our sound data, and a deep neural net classifier. We outline the tradeoffs between the two methods in computational intensity and accuracy, and report further directions for research.

1 Introduction

Gaining meaningful information from an acoustic environment is something that humans do naturally, and so it is a problem that the robotics community also values. Humans split up sounds into two major use cases. The first, is the overall problem of deriving semantic meaning from a sound source. This could include listening to speech and gathering word meaning (as modeled in [1]), or getting information from non-vocal sounds in an environment (such as traffic lights beeping). The second use case for sound classification is musical, and there have been many theories on the biological reasons for these musical cognitive capabilities. A demonstration of one possible musical problem would be trying to discriminate different types of instruments from each other, which is a cognitive ability that may have arose from discrimination of spectral cues in order to communicate with our own species, rather than others.

2 Related Work

In the research community, there is increasing interest in the problem of sound classification, especially pertaining to robotics. To date, a variety of signal processing and machine learning techniques have been applied to this problem, including matrix factorization [2], unsupervised dictionary learning [3], wavelet filterbanks with hidden markov models [4] and more recently deep neural networks [5][6] and deep convolutional neural networks [6]. Deep neural networks are, in particular, very well suited for this problem because they are theoretically able to capture the modulation patterns in time and frequency spectrogram [7].

3 Problem Formulation

Our project aims to explore the problem of sound classification in a generalized sense, and we hope to identify pros and cons of the k-NN and deep neural net classification algorithms in relation to the two sound classification problems that happen in everyday life as mentioned in the introduction. Specifically for this project, we will focus on the musical cognitive ability to discriminate instrument types in an environment. We ask the main question of how well each algorithm can accurately label instrument types, and which algorithm would be better for this task. We believe that our results could then be applied to other problems, such as speaker identification, with some parameter tuning.

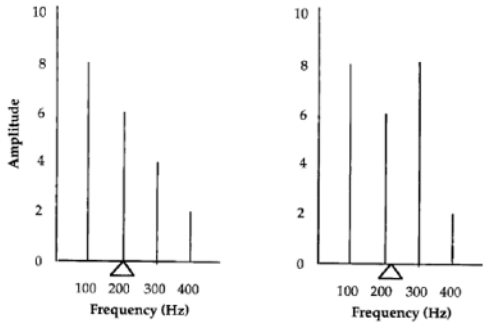
4 Technical Approach

We introduce two classification approaches in machine learning. A simple, k-nearest Neighbors classifier and deep neural networks classifier. Both of these classifiers will use the same set of features that we extract from the audio signal.

4.1 Feature Extraction

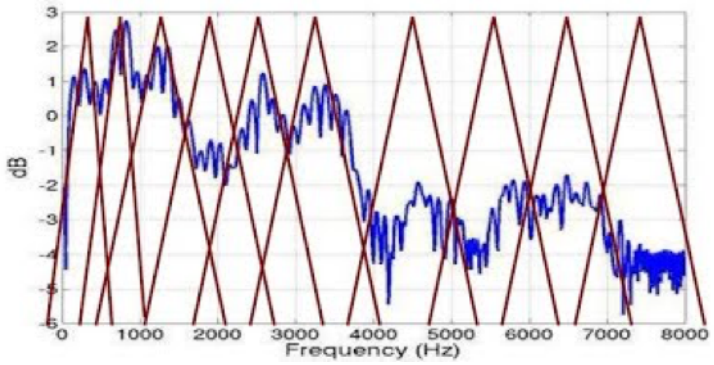
Feature extraction in audio waveforms usually includes gaining useful information in the frequency domain. The most basic example of characterizing a signal by frequencies would be to use the Fourier transform, which reports magnitudes and phases of frequencies of a complex waveform. Initial analysis using Fourier transforms reveals some rough acoustic features of sound. For instance, spectral centroid is a rating that can be used to rate how bright or dark a sound is in general, and the variance can shed light on how high or low the sound is. Spectral flatness is a rating of how noisy the signal is, and in musical context, can be used to measure how percussive a sound is.

Figure 1: Spectral centroids which characterize acoustic brightness



Features derived from Fourier transforms are useful for generalizing types of sounds, but are not the most specific. Mel Frequency Cepstral Coefficients (MFCCs) are based on Fourier transforms, but can give us a great characterization of a sounds behavior over time. Whereas Fourier transforms tell us a great deal about a sound within a certain time slice, the cepstral coefficients are more likely to tell us about frequencies of frequencies over many time slices, and that abstraction helps us characterize change in spectral shape over time, e.i. timbre. In addition, calculating these coefficients on the Mel scale allows for us to get information back that is specifically relevant to human perception of sound. Rather than using a linear scale, the Mel scale is based on judgements of pitch relationships, and so it is more similar to a logarithmic scale. We will calculate around 14 coefficients, and our classifier will try to map these 14 parameters to sound type.

Figure 2: Mel scale filter bank which highlights notable magnitudes of the spectrum

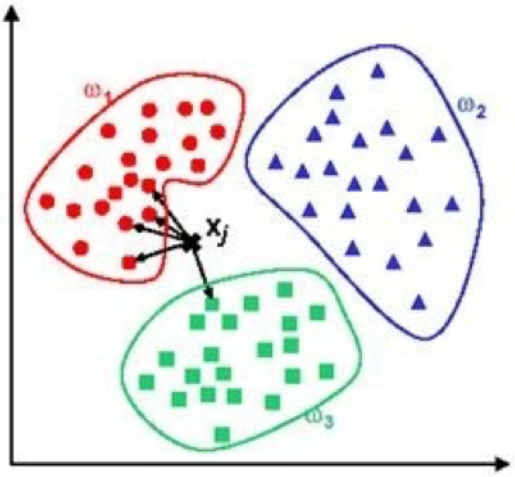


4.2 k-Nearest Neighbors Approach

The k-Nearest Neighbors (k-NN) introduced in [8] employs a voting system that uses euclidian distances to relate new unclassified stimuli to previous categories. First, all of

the provided data is plotted in an N-dimensional space, and the main objective of the search is to find the k-th closest data points to a new input. As an example related to sound, after plotting many piano samples and drum samples in an N-dimensional space (say, based on 14 Mel frequency coefficients), the k-th nearest neighbors to a new drum sound should overwhelmingly vote that the sample is a drum.

Figure 3: Here, a k-NN algorithm is visualized in two dimensions. For the input $\mathbf{X}_j$, 5 nearest neighbors are found and the voting system determines that $\mathbf{X}_j$ should be classified into category 1.

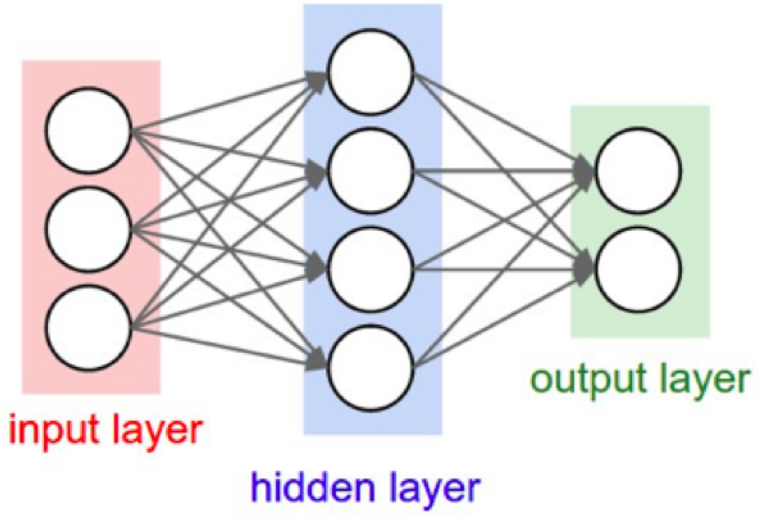


4.3 Deep Neural Networks

Neural networks, or *artificial* neural networks, are a supervised learning approach. It has in recent years gained widespread popularity in machine learning research. To put it simply, it is network that transforms the input layer of data to some output layer of data. For the purposes of our problem, our input data will be audio signals in the frequency domain, and our output data is a label selected from a finite set of possible labels. A neural network is often called *deep* because it consist of as many hidden layers as one want, sandwiched between the input and output layer.

Every element in a layer of data is a node, or otherwise known as *neurons*. Take, for example, a feed forward, densely connected deep neural network of three layers illustrated by the figure below. We see that this is very similar to a directed acyclic graph, with vertices being neurons and edges being weights. This type of neural network is also *feed forward* because it consist of no directed loops or cycles, and it is described as *densely connected* because every neuron in the layer i is involved in the computation of every neuron in layer $i + 1$.

Figure 4: A simple feed forward, densely connected neural network



Let us formally define the computation in a neural network. Denote the i^{th} neuron as $\mathbf{x}_i$, and let $h^l(\cdot)$ be a function that maps a some neuron $\mathbf{x}_i$ to the value of that neuron at layer l. Then we define

$$h^l(\mathbf{x}_i) = f(\mathbf{W}^l h^{l-1}(\mathbf{x}_t) + \mathbf{b}^l) \tag{1}$$

where $f(\cdot)$ is an activation function, $\mathbf{W}^l$ is the weight matrix of layer l, and $\mathbf{b}^l$ is the bias vector of layer l. Some popular choice of activation functions related to sound classification problems are linear ones such as reLu and softMax, or nonlinear ones such as tanh and sigmoid. The bias vector may or may not be necessary, but it is capable of shifting the results to the direction that we intend.

The goal of any machine learning is to iteratively improve our weights until we minimize the cost. For multiclass classification problems, a common cost function choice is cross entropy. We will choose Adam as our optimizer, which is a method of stochastic gradient descent that adaptively decreases the learning rate to avoid overshooting [9].

4.4 Tools and Data

We will use sckit-learn[1] for the k-NN model and Tensorflow[2] for the deep neural network. For feature extraction, we wil use Librosa[3], a python library for audio and music processing. The dataset that we will be working with is prepared by Philharmonia Orchestra[4],

[1] http://scikit-learn.org/stable/
[2] https://www.tensorflow.org
[3] http://librosa.github.io/
[4] http://www.philharmonia.co.uk/explore/sound_samples

which contains sample audio files of a variety of instruments.

5 Expected Results

Our results should indicate how accurately the k-NN and deep learning algorithms were able to identify our test instrument samples. Upon success, we will be able to determine which algorithm is better for classifying instruments. We will weigh accuracy against computational cost, and also acknowledge which algorithms out of the two are feasible in real-time settings.

6 Timeline

- Nov 8: Have a dataset collected, decide which libraries are the best to use.
- Nov 16: Progress report due, possibly have one of the algorithms trained on the dataset.
- Nov 25: Have both algorithms trained.
- Dec 4: Document the success of each algorithm and discuss tradeoffs.

References

[1] McClelland, James L., and Jeffrey L. Elman. "The TRACE model of speech perception." *Cognitive psychology* 18, no. 1 (1986): 1-86.

[2] Mesaros, Annamaria, Toni Heittola, Onur Dikmen, and Tuomas Virtanen. "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations." In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on,* pp. 151-155. IEEE, 2015.

[3] Salamon, Justin, and Juan Pablo Bello. "Unsupervised feature learning for urban sound classification." In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on,* pp. 171-175. IEEE, 2015.

[4] Geiger, Jrgen T., and Karim Helwani. "Improving event detection for audio surveillance using gabor filterbank features." In *Signal Processing Conference (EUSIPCO),* 2015 23rd European, pp. 714-718. IEEE, 2015.

[5] Cakir, Emre, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. "Polyphonic sound event detection using multi label deep neural networks." In *Neural Networks (IJCNN), 2015 International Joint Conference on,* pp. 1-7. IEEE, 2015.

[6] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." *IEEE Signal Processing Letters* 24, no. 3 (2017): 279-283.

[7] Salamon, Justin, and Juan Pablo Bello. "Feature learning with deep scattering for urban sound analysis." In *Signal Processing Conference (EUSIPCO),* 2015 23rd European, pp. 724-728. IEEE, 2015.

[8] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." *IEEE transactions on information theory* 13, no. 1 (1967): 21-27.

[9] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv*:1412.6980 (2014).