A Survey Of Machine Learning For The Detection Of Predominant Musical Instruments

Pratheeksha K S
Courant Institute of Mathematical Sciences
N13908667

December 2016

1 Introduction

Music Information Retrieval(MIR) is the interdisciplinary field of retrieving information from music. It bridges the domains of digital audio signal processing, pattern recognition, software system design, and machine learning. Given recent advances in online music distribution and searching, Music Information Retrieval becomes very relevant for online music distribution and searching. Simply put, MIR algorithms allow a computer to "listen" and "understand or make sense of" audio data, such as MP3s in a personal music collection, live streaming audio, or gigabytes of sound effects, in an effort to reduce the semantic gap between high-level musical information and low-level audio data.

With the identification of musical instruments, extraction of musical sources can be tied to the audio encoding, which can help in copyright verification and similarity searches. A music instrument's sound is characterized by pitch, loudness and timbre. We will study various feature extraction and classification techniques for musical instruments that have been proposed over the time.

2 Basic Approaches

There are 2 basic approaches to identifying music instruments. The first approach focuses on identifying and extracting the features that are indicative of the music instruments present in the audio, and discarding the rest of the audio signals.

Mel Frequency Spestral Coefficients(MFCC) have been used extensively in speech analysis over the past few decades and have more recently received attention in music analysis. Obtaining the MFCCs involves analysing and processing the sound according to the following steps -

1. Divide the signal into frames and get the amplitude spectrum of each frame
2. Take the log of these spectrums and convert to the Mel scale
3. Apply the Discrete Cosine Transform (DCT)

The DCT in step 3 approximates the Principal Component Analysis in that it reduces the data orthnormally, thus leaving a series of uncorrelated values (the coefficients) for each frame of the sound. The first few principal components correspond to the predominant instruments in the audio clip. Hence, this is a natural choice of a feature for predominant music instrument detection.

The second one is a more recent approach, which uses an end to end technique to recognize instruments using raw audio signals. The term end-to-end learning is used to refer to processing architectures where the entire stack, connecting the input to the desired output, is learned from data. An end-to-end learning approach greatly reduces the need for prior knowledge about the problem, and minimises the required engineering effort; only the tuning of the model hyperparameters requires some expertise, but even that process can be automated.

With the contemporary technological advancements in machine learning and deep learning, Convolutional Networks(ConvNets) have become very popular and very successful in classification tasks for images. ConvNet is useful for data with local groups of values that are highly correlated, forming distinctive local characteristics that might appear at different parts of the array. We will explore how this can be extended to musical audio data.

3 Datasets

Below are the datasets that are popularly used in predominant instrument recognition in musical audio.

1. IRMAS: The instruments considered are: cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and human singing voice. This dataset is derived from the one compiled by Ferdinand Fuhrmann in his PhD thesis. The train data contains 6705 audio files in 16 bit stereo wav format sampled at 44.1kHz. They are excerpts of 3 seconds from more than 2000 distinct recordings. Test data has 2874 excerpts in 16 bit stereo wav format sampled at 44.1kHz. The dataset is comprised of annotations for 11 predominant musical instruments.

2. MedleyDB: This was curated primarily to support research on melody extraction, addressing important shortcomings of existing collections. For each song they provide melody f0 annotations as well as instrument activations for evaluating automatic instrument recognition. The dataset is also useful for research on tasks that require access to the individual tracks of a song such as source separation and automatic mixing. This project was lead by Rachel Bittner at NYU's Music and Audio Research Lab, along with Justin Salamon, Mike Tierney and Juan Pablo Bello. The dataset contains 122 Multitracks (mix + processed stems + raw audio + metadata) in 16 bit stereo wav format sampled at 44.1kHz.

4 Notable Results

Here are some of the notable results in the MIR field for instrument recognition.

1. Many of the works in instrument recognition focus on studio recorded isolated notes.

   (a) Eronen et al. used cepstral coefficients and temporal features to classify 30 orchestral instruments and achieved a classification accuracy of 95% for instrument family level and about and about 81% for individual instruments. [5]

   (b) Diment et al. used a modified group delay feature that incorporates phase information together with mel-frequency cepstral coefficients(MFCCs) and achieved a classification accuracy of about 71% for 22 instruments.[4]

   (c) Yu et al. applied sparse coding on cepstrum with temporal sum-pooling and achieved an Fmeasure of about 96% for classifying 50 instruments. [14]

   We will now look at some of the results in the polyphonic music recordings.

2. Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies [6] : This paper focuses on a single music genre (i.e., jazz) but combines a variety of instruments among which are percussion and singing voice. Using a varied database of sound excerpts from commercial recordings, they show that the segmentation of music with respect to the instruments played can be achieved with an average accuracy of 53%.

   (a) MFCCs are extracted from the training data and they comprise of the feature space.

   (b) The dimensionality of the feature space is reduced by principal component analysis (PCA) yielding a smaller set of transformed features to be used for inferring a hierarchical taxonomy.

   (c) A hierarchical clustering algorithm (exploiting robust probabilistic distances between possible classes) is used to generate the targeted taxonomy.

   (d) SVM classifiers are trained on these features for every node of the taxonomy on a "one versus one" basis.

3. Deep convolutional neural networks for predominant instrument recognition in polyphonic music [8]: This paper makes use of ConvNets to identify predominant instruments in music.

   (a) This paper makes use of the IRMAS dataset. In the first preprocessing step, the stereo input audio is converted to mono by taking the mean of the left and right channels, and then it is downsampled to 22,050 Hz from the original 44,100 Hz of sampling frequency. In the next step, the MFCCs are extracted.

   (b) A ConvNet architecture similar to VGG is used to train the network.

   (c) The paper also discusses various activation functions(ReLU/ LReLU/ tanh) and achieves 65% accuracy in instrument detection in polyphonic music .

4. Automatic Instrument Recognition

   (a) To compare end-to-end learning with the traditional MIR approach of using a mid-level representation of the audio signals, Sander Dieleman et al. trained deep CNNs to perform automatic tagging on the Magnatagatune dataset. [11]

   (b) Peter Li et al. showed that an end-to-end approach by learning directly from raw audio signals, thus further reducing the amount of prior knowledge required, on the MedleyDB dataset produced a higher accuracy.

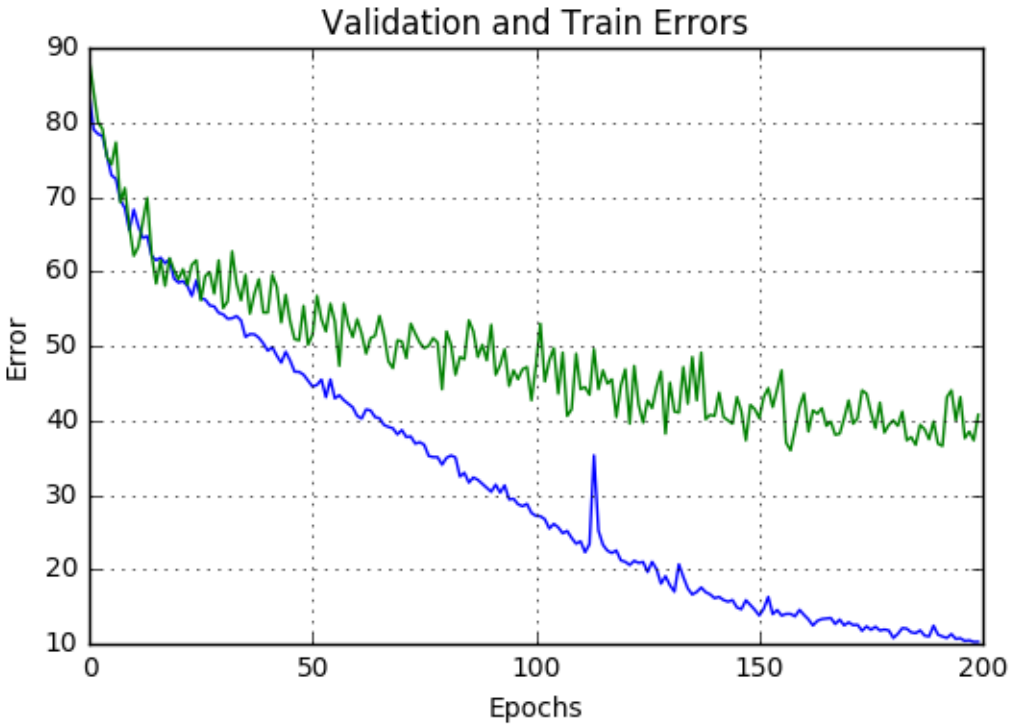

Figure 1: Cross-validation performance over 200 epochs

5 Experiments

We implemented the instrument recognition techniques described in Deep convolutional neural networks for predominant instrument recognition in polyphonic music [8] and experimented with some of the parameters. This paper uses the IRMAS dataset. [2]

1. Data preprocessing: As described in the paper, we extract the MFCCs for each of the audio clips in the training data using Librosa, a python module that enables audio processing.

2. Architecture: A simple ConvNet architecture has the following components - [INPUT - CONV - RELU - POOL - FC]

   - INPUT: In our case, MFCC constitutes the input, which is a [20x130] Tensor.
   - CONV: This layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume.
   - RELU: This layer will apply an element-wise activation function, such as the max(0,x)(thresholding at zero). This leaves the size of the volume unchanged.
   - POOL: This layer will perform a downsampling operation along the spatial dimensions (width, height).
   - FC: The fully-connected layer will compute the class scores, resulting in volume of size [1x1x11], where each of the 11 numbers correspond to a class score, such as among the 11 categories of musical instruments. As with ordinary Neural Networks and as the name implies, each neuron in this layer will be connected to all the numbers in the previous volume.

   We also added batch normalizations and dropouts that help in regularizing the training across minibatches and prevent overfitting, respectively.

3. Results: We used ReLUs as activation functions. We experimented with both stochastic gradient descent and Adam for optimization functions. We used a learning rate of .01 with .0001 learning rate decay and .001 weight decay as SGD parameters.

   The model performed on par with the baseline, with 36% cross-validation error (minimum) when run for 200 epochs on a GeForce GTX 1080 GPU.

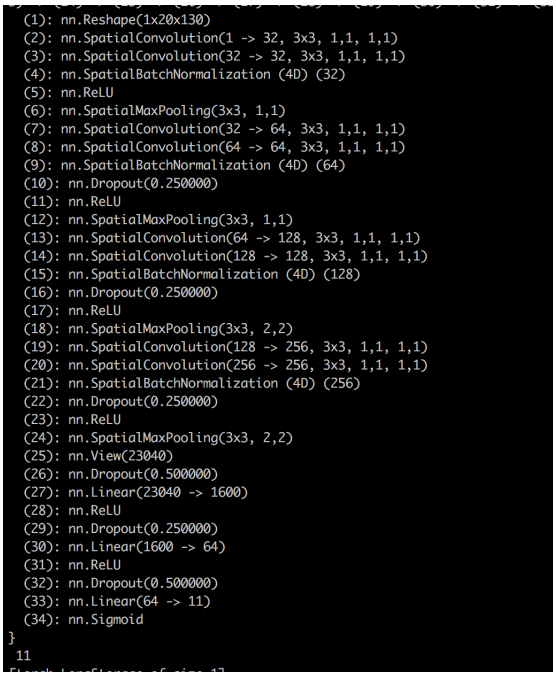

Figure 2: Full model used for learning musical instruments in IRMAS dataset

6 Conclusion and Future Work

Music Information Retrieval is an exciting field of research with various machine learning opportunities. Instrument recognition has several applications. It can also help us understand the source of music better and hence, a better understanding of this music information can help improve audio streaming and recognizing platforms.

Using ConvNet to learn music information is a fairly new research topic in MIR. We can engineer a network that focuses on music data, and not just look at it as an extension of the image classification problem. Audio data along with MFCC can be used to train the ConvNet. Generalizing these models for music data collected from varied sources, in the presence of noise, is also something we can explore. We can explore the newer and more powerful models like Residual Neural Networks, along with varied initializations and activation functions.

References

[1] R. Bittner, J. Salamon, M. Tierney, C. Cannam M. Mauch, and J. P. Bello. *MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.* ISMIR, 2014.

[2] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera. *A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals.* ISMIR, 2012.

[3] Ronan Collobert, Laurens van der Maaten, and Armand Joulin. *Torchnet: An Open-Source Platform for (Deep) Learning Research.* FAIR, 2016.

[4] A. Diment, P. Rajan, T. Heittola, and T. Virtanen. *Modified group delay feature for musical instrument recognition.* IEEE, 2013.

[5] A. Eronen and A. Klapuri. *Musical instrument recognition using cepstral coefficients and temporal features.* IEEE, 2000.

[6] Slim Essid, Gaël Richard, and Bertrand David. *Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies.* IEEE, 2006.

[7] Glenn Eric Hall, Hassan Ezzaidi, and Mohammed Bahoura. *Advanced Machine Learning Technologies and Applications.* Springer Berlin Heidelberg, 2014.

[8] Yoonchang Han, Jaehun Kim, , and S Kyogu Lee. *Deep convolutional neural networks for predominant instrument recognition in polyphonic music.* IEEE, 2016.

[9] Peter Li, Jiyuan Qian, and Tian Wang. *Automatic Instrument Recognition In Polyphonic Music Using Convolutional Neural Networks*. 2015.

[10] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. *Rectifier Nonlinearities Improve Neural Network Acoustic Models*. ICML, 2013.

[11] City University of London. *MagnaTagTune dataset*. 2009.

[12] Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. ICLR, 2015.

[13] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, and Yao Zhao Shuicheng Yan. *The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification*. IEEE, 2014.

[14] L.-F. Yu, L. Su, and Y.-H. Yang. *Sparse cepstral codes and power scale for instrument identification*. ICASSP, 2014.