Dhvani - A Deep Learning Approach for Urban Sound Classification using Multilayer Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks.

*

Noopur Rajesh Kumar Kalawatia
dept. of Computer Science and Engineering
University of Florida
Gainesville, United States of America
noopur.rajeshkum@ufl.edu

Ruchika Mishra
dept. of Computer Science and Engineering
University of Florida
Gainesville, United States of America
ruchika.mishra@ufl.edu

***Abstract*—Sound classification is increasingly becoming one of the areas of prime focus for scientific research. Up until now automatic speech recognition has dominated the landscape of sound classification, with newer techniques from deep learning, we can arrive at more sophisticated solutions for the problems encountered in sound classification. Sound classification is an effective tool for environment sound classfication and content based media retrieval systems. In this paper we discuss about the various techniques of deep learning used for urban sound classification like Multilayer Neural Networks, Convolutional Neural Networks and Recurrent Neural Networks. We discuss the methods we have used to calculate the various important features for our model.We go on further detail to arrive at the effectiveness of every model by comparing the performance of the models and choose the best model for the classification.**

***Index Terms*—Urban Sound,Mel Banks Cepstral Coefficients, Tonnetz, Spectogram, Chroma STFT, Feature extraction, Neural networks.**

I. Introduction

Automatic sound classification finds applications in the field of environmental studies for easy retrieval of data and large scale multimedia indexing. Environmental sound retrieval comprises all types of sound that are neither speech nor music. Since this domain is arbitrary in size, most investigations are restricted to a limited domain of sounds.

A survey of techniques for feature extraction and classification in the context of environmental sounds is given in [4]. One of the examples in industries is the company Swinetech. The developers at Swinetech, are trying to develop technology that can prevent prevent piglet deaths due to crushing by their sows, a big problem for hog farmers.Currently speech recognition dominates the landscape of sound classification as researchers often experience difficulty to extract the appropriate features from sound wave forms.

Another problem faced by the researchers is the lack of annotated data to develop appropriate models for classification. The dataset that we have considered for our project is the UrbanSound8K dataset provided by Justin Salamon, Christopher Jacob and Juan Pablo Bello. The names of the audio files contain various meta-data of which we will use the class id, in other words, the label of each audio recording is contained in the file name. The recordings are conveniently pre-sorted into 10 folds to help with the reproduction and comparison of results.The difficulty with sounds is that it cannot be conveniently transformed into vector since the sound waveform is riddled with a lot of unnecessary information. In general, features should capture audio properties that show high variation across the available (classes of) audio objects. It is not reasonable to extract features that capture invariant properties of the audio objects, since they do not produce discriminatory information.[2]

The various features that we are using for our models are as follows,

- MFCC: Mel-frequency cepstral coefficients.
- Melspectrogram: Acoustic time-frequency representation of sound
- Chorma-stft: Entire spectrum is divided into 12 bins representing 12 semitones of the musical octave.
- Spectral contrast: Decibel difference between peaks and valleys in the spectrum.
- Tonnetz : It is a conceptual lattice diagram representing tonal space

For the Recurrent and Convolutional neural networks, we need to preserve the time series format of the raw data to perform classification. To accomplish this task we apply the filterbank and log filterbank techniques to extract the appropriate features from our sound dataset. The method of extracting the same is mentioned in [1].

II. Related Work

There have been various different classifiers to build efficient classifier model to address urban sound classification. The methods and accuracy computed in [4] serves as the baseline for our project. The main conclusions from the project includes,

- The dataset was engineered with required preprocessing to transform each waveform into equal length. After this transformation, the model that worked the best for classification was the deep neural network.
- There is visible overfitting issue in the case of Recurrent Neural Network as the data samples is of a small size.
- Complex neural network models tend to have decreased accuracy in comparision to the simpler multilayer neural network.

III. System Architecture

A typical Sound Classifier consists of modules like input module, feature extractor module and the classifier that classifies the input in the required categories. Our system model is also based on this architecture.

A. Input Module

The input module corresponds to the dataset. For our models, we use the UltraSound8k dataset. This dataset was made by Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. Further explanation is provided in the section IV-A.

B. Feature Extractor

The feature extractor module extracts the required features of the audio waveform.We will be using two sets of datasets for the models, namely MFCCS with other supporting features. The second dataset is the filterbanks and log filterbanks. The technique of extraction for features is provided in the section IV-B.

C. Classifier Module

The classifier module consists of Multilayer Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks.

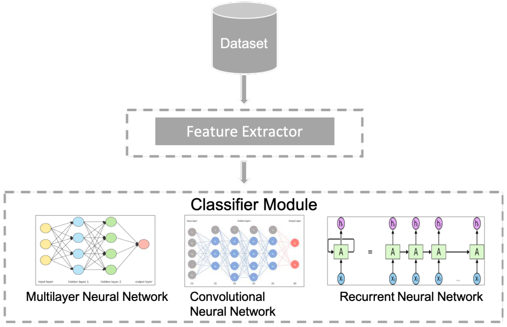

Fig. 1. System Model

IV. System Design and Implementation

A. Dataset

We have used the Ultrasound8K dataset made by Justin Salamon, Christopher Jacob and Juan Pablo Bello. This dataset contains 8732 labeled sound excerpts (¡=4s) of urban sounds from 10 classes: Air conditioner, Children playing, Dog barking, Siren, Engine idling,Gun shot, Jackhammer, Drilling, Car horn, Street horn. The classes are drawn from the urban sound taxonomy described in [2].The distribution in the classes of the dataset is as shown in the Figure-1.

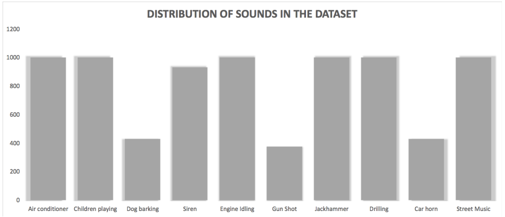

Fig. 2. Example of a figure caption.

The dataset is divided into ten distinct categories with an even distribution. Thus the data if selected randomly will not give over optimistic results on classification. This always solves the possible complications with folding of the data in training and testing datasets.

B. Feature Extraction

From the dataset it was evident that even the waveform with the shortest span of audio had close to 44000 values resulting in an expensive cost to process. As large number of data points translates to equally large number of input units for neural network, it is necessary to extract a smaller number of precise features to minimize the cost of the model. Thus it is necessary to extract precise features that give an accurate representation of the raw data to build efficient classifiers. The next step of contention is to choose the best content based representation for the features, there is an exhaustive list to choose from.

One of the most prominent features often employed for sound classification is Mel-frequency cepstral coefficients (MFCCs). MFCC is defined as a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc.[3]

MFCC are calculated as follows, the method described below is synthesized from *Mel Frequency Cepstral Coefficient (MFCC) tutorial by James Lyons*

- Based on the assumption that for short durations, sound remains statistically stagnant, we reduce the frame size to 20-40 frames.
- Calculation of power spectogram of each frame. Power spectogram gives an indication of frequencies present in the frames.

- The periodogram spectral estimate consists of a lot of frequencies which do not provide necessary information about the frames. Mel filterbanks further synthesis the information into bins inorder to understand the frequency distribution, thus arriving at the energy distribution of the frames.
- Calculate the logarithm of the filterbank energies
- Compute the Discrete Cosine Transform(DCT) of the log filterbank energies.
- The MFCCs are the amplitudes of the resulting spectrum.

The other features that can be extracted from the raw data are Melspectogram, Chroma-stft, Spectral Contrast, Tonnetz.The Librosa library has the provosion of providing all the above features from raw data. Please refer to figure 2 for observing the raw data and Figure 2 and Figure 3 for visualization of the Log Power spectogram and Power Spectogram of the raw data.

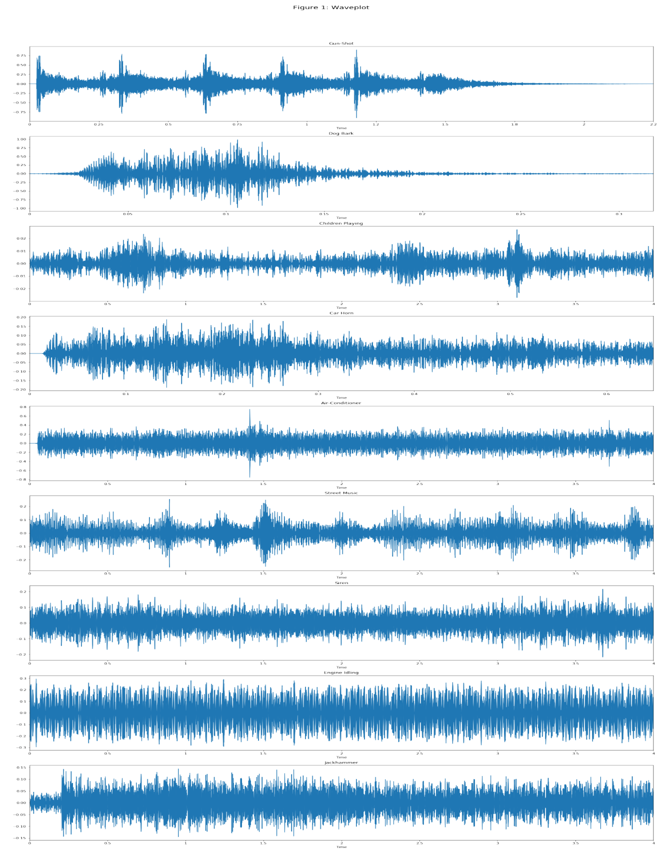

Fig. 3. Raw data for waveforms from the samples of every class

The main motivation to use the second set of features is, the filterbanks computed in the above sequence of step are highly correlated, which could be problematic in some machine learning algorithms.The second set of features that we will be computing are the filter banks for our sound waveforms. The main distinction between filter banks and MFCC is, filter banks are used if the machine learning algorithm is not susceptible to highly correlated input. We use MFCCs if the machine learning algorithm is susceptible to correlated input.

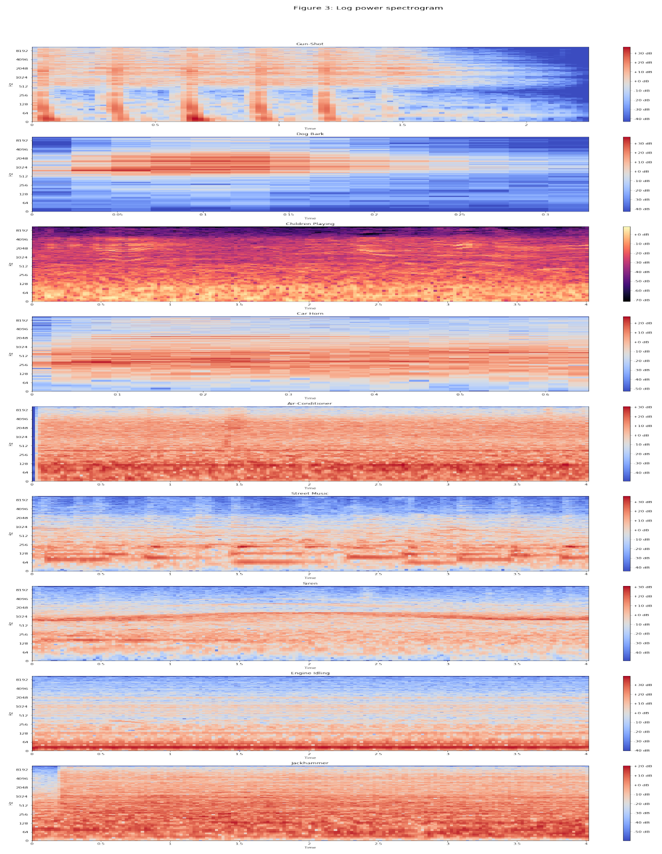

Fig. 4. Log Power Spectograms

One of the important issues faced by us while computing the filter banks was the varying lengths of the audio data. It is imperative to have the input data be of the same length to maintain the integrity of the input data for the models. To keep the data of the same length we employ two techniques, the zero padding and maintaining the frame size same for all the input wave forms.

Another important to consider using MFCC is, they were predominantly useful while using Gaussian Mixture Models - Hidden Markov Models. MFCCs evolved as the standard features for automatic speech recognition. Deep Learning models are less susceptible to highly correlated input and therefore the MFCC are not highly favorable. It is beneficial to note that Discrete Cosine Transform (DCT) is a linear transformation, and therefore undesirable as it discards some information in speech signals which are highly non-linear and probably useful.

C. *Classifier Module*

- **Model 1 : Multilayer Neural Network** A multilayer neural network can be defined as a class of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called

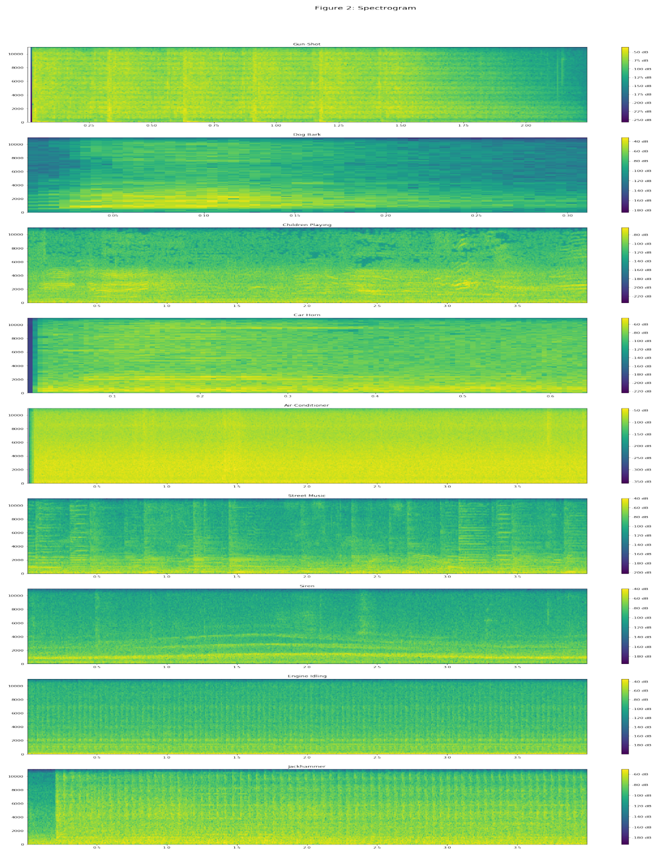

Fig. 5. Power Spectograms

backpropagation for training. Currently we are working on improving the network and understanding the issues of overfitting, learning rate and activation functions that can further optimize our results.

Model 2 : Convolutional Neural Network CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing.They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

Model 3 : Recurrent Neural Network A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence.RNNs can use their internal state (memory) to process sequences of inputs.

V. Performance Evaluation

Currently we have completed the multilayer neural network model. The multilayer neural network that we trained has three hidden layers and has 280, 300, 270 neurons respectively. We have achieved a test accuracy of about 0.893 for our dataset. The model was run on i5 8GB RAM processor. The experimental results of the same are presented in the figure - 6

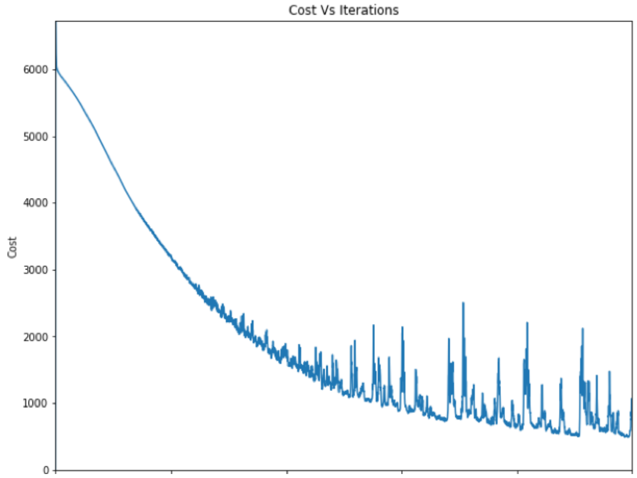

Fig. 6. Results from MLP

VI. Acknowledgements

Our thanks to Justin Salamon, Christopher Jacoby, and Juan Pablo Bello for creating the UrbanSound8K dataset.

References

[1] Dalibor Mitrovic, Matthias Zeppelzauer, Christian Breiteneder. Features for Content-Based Audio Retrieval

[2] Shuhui Qu, Juncheng Li, Wei Dai, Samarjit Das, LEARNING FILTER BANKS USING DEEP LEARNING FOR ACOUSTIC SIGNALS in arXiv:1611.09526v1 [cs.SD] 29 Nov 2016

[3] J. Salamon, C. Jacoby and J. P. Bello, ”A Dataset and Taxonomy for Urban Sound Research”, 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.

[4] Meinard Mller (2007). Information Retrieval for Music and Motion. Springer. p. 65. ISBN 978-3-540-74047-6.

[5] Chih-Wei Chang, Benjamin Dorman. Urban Sound Classification: With Random Forest, SVM, DNN, RNN, and CNN Classifiers