Urban Sound Classification

Qurat ul Aain, Zara Akhtar
Istanbul Sehir University
34865 Dragos, Istanbul, Turkey
{qurataain,zaraakhtar2019}@std.sehir.edu.tr

Abstract

Sound classification is a burgeoning field which has been growing in parallel to developments in machine learning and urban informatics. In this project, we define methodology of using Neural Network and Convolutional Neural Network to perform multi-class classification of an urban sound dataset of environmental sounds. This paper evaluates the accuracy of performing these techniques on classifying our dataset by extracting features i.e. components of audio signal that are good for identifying linguistic content. Mel Frequency Cepstral Coefficient or MFCCs are a feature widely used in sound classification. MFCCs are a small set of features (40 in our project) which concisely describe the overall shape of a spectral envelope.

The accuracy of the network is evaluated on public dataset of environmental and urban recordings. The model outperforms baseline implementations relying on mel-frequency cepstral coefficients (MFCCs) and achieves results comparable to other state-of-the-art approaches.

Dropout regularization is implemented to prevent the neural network from overfitting the data, and to improve the regularization.

1 Introduction

Convolutional neural networks date back as far as to the 1980s[1], yet only recently have they been adopted as a method of choice for various object classification tasks.

Although primarily used in visual recognition contexts, convolutional architectures have been also successfully applied in speech [2] and music analysis [3]. At the same time, classification of environmental sounds is still predominantly based on applying general classifiers (Gaussian mixture models, support vector machines, hidden Markov models) to manually extracted features, such as melfrequency cepstral coefficients (MFCCs).

This contrast in the development of different fields raises a research question which has not yet been widely addressed. Can convolutional neural networks and simple neural networks be effectively used in classifying environmental and urban sound sources? The goal of this paper is to show how this can be done.

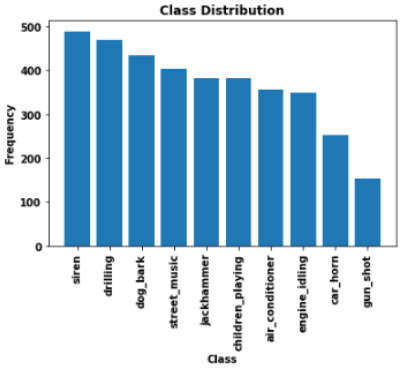

Figure 1: Count of items in each class.

In essence, convolutional neural networks are a simple extension of the multilayer perceptron model. However, their architectural differences have significant practical consequences.

A typical convolutional neural network consists of a number of different layers stacked together in a deep architecture: an input layer, a group of convolutional and pooling layers (which can be combined in various ways), a limited number of fully connected hidden layers, and an output (loss) layer. The actual difference, when compared to the multilayer perceptron, lies in the introduction of a combination of convolution and pooling operations.

The dataset contains 8732 sound excerpts (¡=4s) of urban sounds from 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. Figure 2 shows the count of items in each class.

When given an audio sample in a computer readable format (such as a .wav file) of a few seconds duration, we want to be able to determine if it contains one of the target urban sounds with a corresponding Classification Accuracy score. Each sample is the amplitude of the wave at a particular time interval.

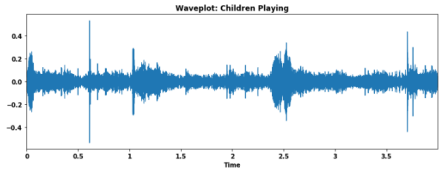

Figure 2: Waveplot of Children Playing

1.1 Sound Classification

One of the main problems with training deep neural architectures in a supervised manner is the amount of computational effort and labeled data required for efficient learning. While the former is in some part addressed on a universal basis by hardware advances and general-purpose GPU computing, the latter is very domain-dependent.

Figure 1 shows the waveplot of children playing.

1.2 Motivation

Urban noise recognition has attracted a lot of attention in the recent years in city management and safety operations, especially in the recent smart city engineering. Environmental sound classification is a growing area of research with numerous real world applications. Whilst there is a large body of research in related audio fields such as speech and music, work on the classification of environmental sounds is comparatively scarce.

Likewise, observing the recent advancements in the field of image classification where convolutional neural networks are used to to classify images with high accuracy and at scale, it begs the question of the applicability of these techniques in other domains, such as sound classification.

Some of the practical uses of sound classification are:

i. Assisting deaf individuals in their daily activities

ii. Smart home use cases such as 360-degree safety and security capabilities

iii. Industrial uses such as predictive maintenance

iv. Content-based multimedia indexing and retrieval

1.3 Challenges

The main challenge in our project was handling the 2000 or so corrupt files or missing or mismatched indexes. Using a function parser that we created, we were able to identify the files that were either corrupt their IDs were mismatched. We created csv files of those indexes for both test and train set. we were able to mask those indexes in the original train and test csv files. Therefore, our code does not read those files.

2 Literature Review

A neural network model for a mechanism of visual pattern recognition is proposed in this paper [1].

Urban noise recognition play a vital role in city management and safety operation, especially in the recent smart city engineering [4].

Convolutional deep belief networks are applied to audio data and empirically evaluated them on various audio classification tasks [2].

Convolutional Neural Network were successfully applied to audio data of music files [3].

We are implementing Max Pooling using Keras Layers to regularize our model [5].

ReLU and Softmax Activation Functions are used. This is the default recommendation in modern neural networks [6].

Dropout regularization is used to reduce overfitting the data and improve generalization in Neural Networks [7].

We are implementing Max Pooling using Keras Layers to regularize our model [5].

3 Proposed Methodology

The focus of our project is the categorization of urban environmental sounds using machine learning techniques. We are doing multi-class classification using Neural Networks and Convolutional Neural Networks in this project.

A novel methodology that we are using here and that can be pursued further in the future is using Simple and Convolutional Neural Networks for sound classification.

The first step in any sound classification is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise.

The feature extraction from sound classification is done using Neural Networks and Convolutional Neural Networks.

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in sound classification. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. The mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10-20) which concisely describe the overall shape of a spectral envelope.

By printing the shape of mfccs you get how many mfccs are calculated on how many frames. The first value represents the number of mfccs calculated and another value represents a number of frames available.

Figure 3 shows the MFCC plot and Figure 4 shows the spectrogram view of the MFCC plot.

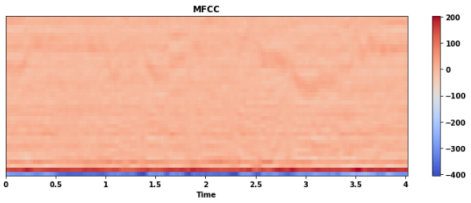

Figure 3: MFCC plot

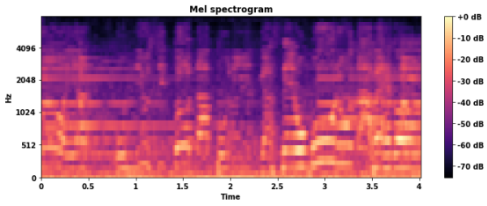

Figure 4: Spectrogram of MFCC plot

Equation (1) shows the formula for converting from frequency to Mel scale is:

$$M(f) = 1125ln(1 + f/700) \quad (1)$$

Since we are performing multi-class classification, and our activation function requires a multi-class output, we cannot use tanh or Sigmoid functions, since they saturate around +1 and -1. Further, the functions are only really sensitive to changes around their mid-point of their input, such as 0.5 for Sigmoid and 0.0 for tanh. Therefore, we use ReLU and Softmax Activation Functions. This is the default recommendation in modern neural networks [6]. We use Softmax as the output function of the last layer in neural networks (if the network has n layers, the n-th layer is the Softmax function). This fact is important because the purpose of the last layer is to turn the score produced by the neural network into values that can be interpreted by humans.

Figure 5 and Figure 6 show the plot of training and validation accuracy and loss per epoch (total 100 epochs used) for Neural Networks and for CNN respectively.

In a neural network, the activation function is responsible for transforming the summed weighted input from the n ode into the activation of the node or output for that input.

Dropout regularization is implemented to prevent the neural network from overfitting the data, and to improve the regularization.

An automatic categorization process takes a segment of audio, and returns the “classes” or “labels” present in the segment.

We are implementing Max Pooling using Keras Layers to regularize our model. It contains four layers, the sequential layer, and the two hidden layers use ReLU as the

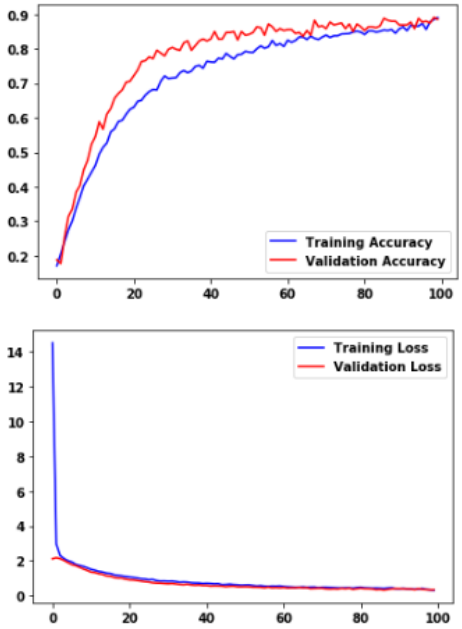

Figure 5: Training and Validation Accuracy / Loss per epoch for Neural Networks

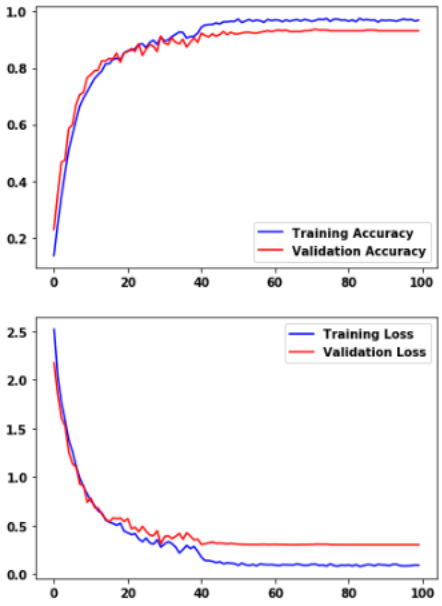

Figure 6: Training and Validation Accuracy / Loss per epoch for Convolutional Neural Networks

	Classification Accuracy	
Network Type	$Train$	$Validation$
Neural Network	87%	90%
CNN	97%	92%

Table 1: Classification Accuracy of Simple Neural Network and Convolutional Neural Network

activation function, while output layer uses Softmax as the activation function. Max Pooling is a pooling operation that calculates the maximum, or largest, value in each patch of each feature map. In a nutshell, the reason is that features tend to encode the spatial presence of some pattern or concept over the different tiles of the feature map (hence, the term feature map), and it's more informative to look at the maximal presence of different features than at their average presence [5].

Table 1 summarizes the performance of the Neural Network and Convolutional Neural Networks performance.

4 Conclusions

We defined a methodology of using Neural Network and Convolutional Neural Network to perform multi-class classification on urban sound dataset. We found classification accuracy of 87% to 91% on validation set using Neural Network and accuracy of 92% to 97% using CNN. This is higher than we initially expected.

Our paper used Mel Frequency Cepstral Coefficient of MFCCs to extract features for sound classification. This is the current state-of-the art on this emerging field of urban sound classification.

References

[1] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," 1980.

[2] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," 2009.

[3] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR, 2011.

[4] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.-P. Vidalütze, "Urban noise recognition with convolutional neuralnetwork," 2018.

[5] F. Chollet, *Deep Learning with Python*, vol. 1. Manning Publications Co, 2017.

[6] S. Alvarez and B. Puente, "Single and multi-label environmental sound classification using convolutional neural networks," 2018.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overtting," 2014.