Classifying Urban sounds using Deep Learning

Students :

Nasser Alshehri

Abdulrahman Alqurashi

20 January, 2022

1. Introduction

Sounds are all around us. Whether directly or indirectly, we are always in contact with audio data. Sounds outline the context of our daily activities, ranging from the conversations we have when interacting with people, the music we listen to, and all the other environmental sounds that we hear on a daily basis such as a car driving past, the patter of rain, or any other kind of background noise.

Automatic 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, where discrete sounds happen over time.

The goal of this capstone project, is to apply Deep Learning techniques to the classification of environmental sounds, specifically focusing on the identification of particular urban sounds.

There is a plethora of real world applications for this research, such as:

Assisting deaf individuals in their daily activities

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

Automotive where recognising sounds both inside and outside of the car can improve safety

2. Problem Statement

The objective of this project will be to use Deep Learning techniques to classify urban sounds.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 likelihood score. Conversely, if none of the target sounds were detected, we will be presented with an unknown score.

To achieve this, we plan on using different neural network architectures such as Multi-Layer Perceptronss (MLPs) and Convolutional Neural Networks (CNNs).

3. Data Description

UrbanSound dataset

will use a dataset called Urbansound8K. The dataset contains 8732 sound ex- cerpts (<=4s) of urban sounds from 10 classes, which are:

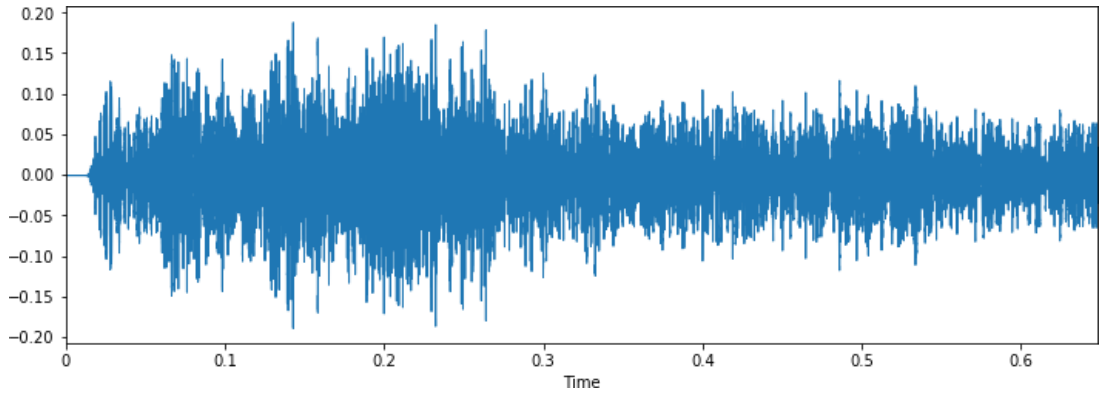
- Air Conditioner
- Car Horn
- Children Playing
- Dog bark
- Drilling
- Engine Idling
- Gun Shot
- Jackhammer
- Siren
- Street Music

These sound excerpts are digital audio files in .wav format.

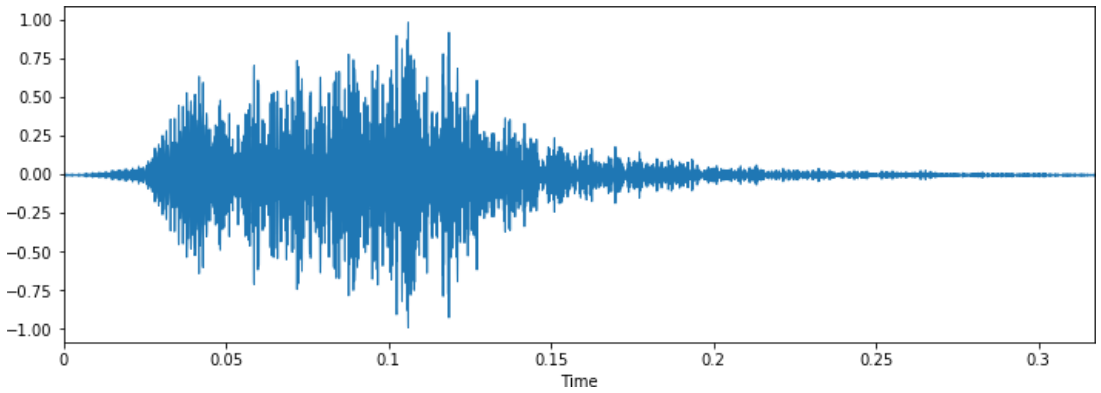
4. Visual inspection

We will load a sample from each class and visually inspect the data for any patterns. We will use librosa to load the audio file into an array then librosa.display and matplotlib to display the waveform .

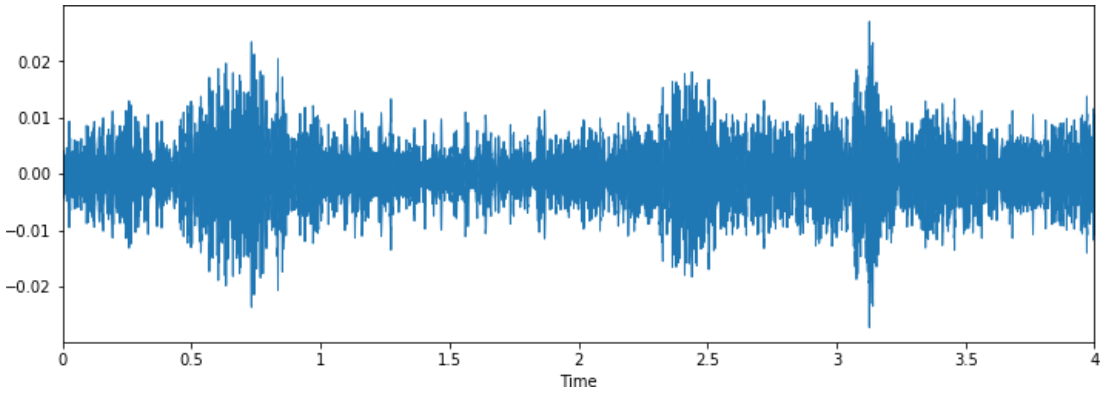
Class: Car horn



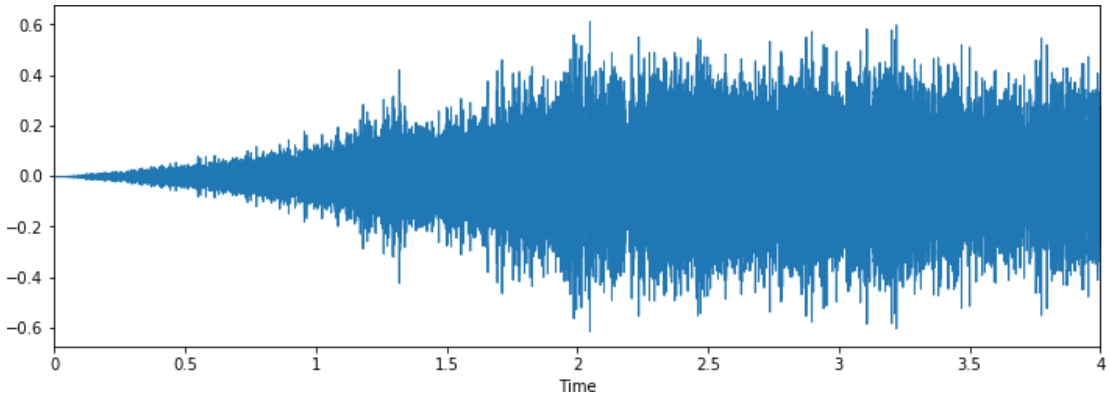
Class: Children playing



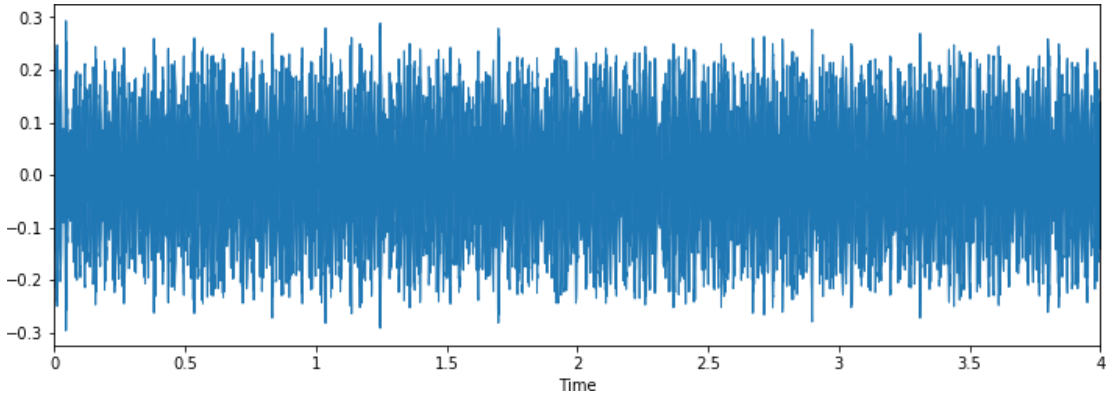
Class: Dog bark



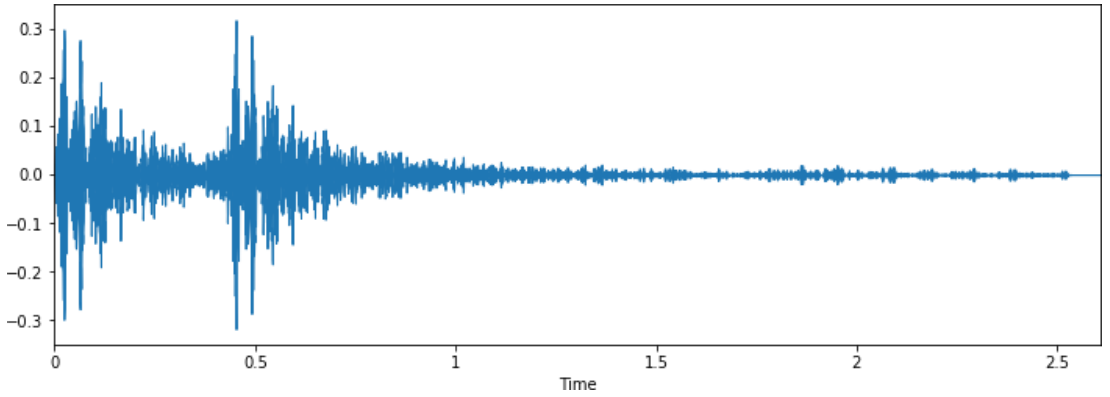
Class: Drilling



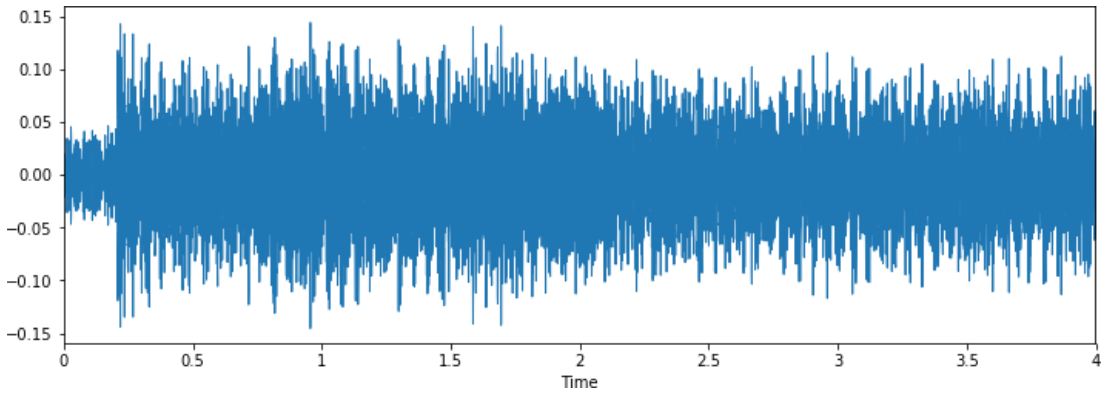
Class: Engine Idling



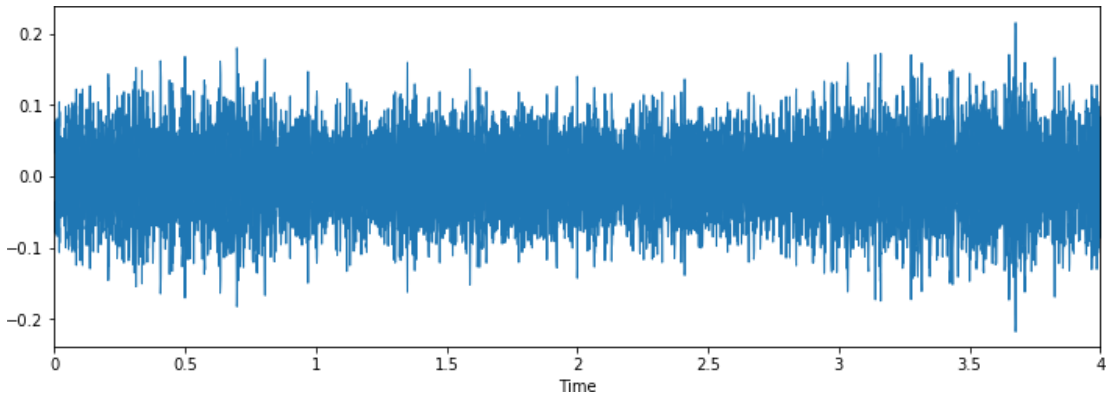
Class: Gunshot



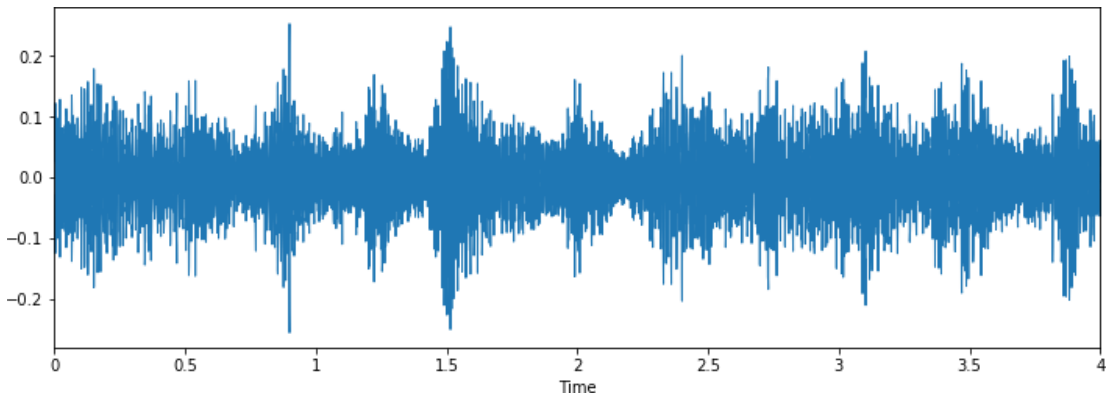
Class: Jackhammer



Class: Siren



Class: Street music



Observations

From a visual inspection we can see that it is tricky to visualise the difference between some of the classes.
Particularly, the waveforms for repetitive sounds for air conditioner, drilling, engine idling and jackhammer are similar in shape.
Likewise the peak in the dog barking sample is similar in shape to the gun shot sample (albeit the samples differ in that there are two peaks for two gunshots compared to the one peak for one dog bark). Also, the car horn is similar too. There are also similarities between the children playing and street music.

5. Model

Initial model architecture - MLP

We will start with constructing a Multilayer Perceptron (MLP) Neural Network using Keras and a Tensorflow backend.

Starting with a sequential model so we can build the model layer by layer.

We will begin with a simple model architecture, consisting of three layers, an input layer, a hidden layer and an output layer. All three layers will be of the dense layer type which is a standard layer type that is used in many cases for neural networks.

Test the model

The accuracy of the model on both the training and test data sets.

Training Accuracy: 0.9252684323550465
Testing Accuracy: 0.8763594734511787

The initial Training and Testing accuracy scores are quite high. As there is not a great difference between the Training and Test scores (~5%) this suggests that the model has not suffered from overfitting.

Predictions

Here we will build a method which will allow us to test the models predictions on a specified audio .wav file.

Validation

Test with sample data Initial sanity check to verify the predictions using a subsection of the sample audio files we explored in the first notebook. We expect the bulk of these to be classified correctly.

Refinement

In our initial attempt, we were able to achieve a Classification Accuracy score of:

Training data Accuracy: 92.3%

Testing data Accuracy: 87%

We will now see if we can improve upon that score using a Convolutional Neural Network (CNN).

(CNN) model

We will modify our model to be a Convolutional Neural Network (CNN) again using Keras and a Tensorflow backend.

Again we will use a sequential model, starting with a simple model architecture, consisting of four Conv2D convolution layers, with our final output layer being a dense layer.

Test the model

Here we will review the accuracy of the model on both the training and test data sets.

Training Accuracy: 0.9819613457408733

Testing Accuracy: 0.9192902116210514

The Training and Testing accuracy scores are both high and an increase on our initial model. Training accuracy has increased by ~6% and Testing accuracy has increased by ~4%.

Predictions

Here we will modify our previous method for testing the models predictions on a specified audio .wav file.

Validation

Test with sample data As before we will verify the predictions using a subsection of the sample audio files we explored in the first notebook. We expect the bulk of these to be classified correctly.

6. Conclusion

It was noted in our data exploration, that it is difficult to visualise the difference between some of the classes.

- Repetitive sounds for air conditioner, drilling, engine idling and jackhammer.
- Sharp peaks for dog barking and gun shot.
- Similar pattern for children playing and street music .