Udacity Machine Learning Nanodegree 2018

Capstone Proposal

Classifying Urban sounds using Deep Learning

Mike Smales

November 2018

1 Domain Background

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. The human brain is continuously processing and understanding this audio data, either consciously or subconsciously, giving us information about the environment around us.

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:
- Content-based multimedia indexing and retrieval
- 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
- Industrial uses such as predictive maintenance

My personal motivation for working on sound classification is my background in DSP and Audio processing. Having worked on a number of projects in this field over the years, most recently at audio connectivity startup chirp.io, I am keen to apply my machine learning knowledge to this domain.

2 Problem Statement

The main 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.

3 Datasets and Inputs

For this project we will use a dataset called Urbansound8K [1]. The dataset contains 8732 sound excerpts (<=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

The accompanying metadata contains a unique ID for each sound excerpt along with it's given class name.

These sound excerpts are digital audio files in .wav format. Sound waves are digitised by sampling them at discrete intervals known as the sampling rate (typically 44.1kHz for CD quality audio meaning samples are taken 44,100 times per second). Each sample is the amplitude of the wave at a particular time interval, where the bit depth determines how detailed the sample will be also known as the dynamic range of the signal (typically 16bit which means a sample can range from 65,536 amplitude values). Therefore, the data we will be analysing for each sound excerpts is essentially a one dimensional array or vector of amplitude values.

4 Solution Statement

The proposed solution to this problem is to apply Deep Learning techniques that have proved to be highly successful in the field of image classification.

First we will extract Mel-Frequency Cepstral Coefficients (MFCC) [2] from the the audio samples on a per-frame basis with a window size of a few milliseconds. The MFCC summarises the frequency distribution across the window size, so it is possible to analyse both the frequency

and time characteristics of the sound. These audio representations will allow us to identify features for classification.

The next step will be to train a Deep Neural Network with these data sets and make predictions. I believe that this will be very effective at finding patterns within the MFCC's much like they are effective at finding patterns within images.

We will use the evaluation metrics described in later sections to compare the performance of these solutions against the benchmark models in the next section.

5 Benchmark Model

For the benchmark model, we will use the algorithms outlined in the paper "*A Dataset and Taxonomy for Urban Sound Research*" (Salamon, 2014) [3]. The paper describes five different algorithms with the following accuracies for a audio slice maximum duration of 4 seconds.

Algorithm	**Accuracy**
SVM_rbf	68%
RandomForest500	66%
IBk5	55%
J48	48%
ZeroR	10%

6 Evaluation Metrics

The evaluation metric for this problem is simply the Accuracy Score.

7 Project Design

Data Preprocessing

First identify the different data types in our dataset and what preprocessing needs to be done to make it uniform.

- resample so all audio had the same sample rate and bit depth

- make sure the sample duration is uniform
- Consider any data augmentations, such as adding background noise (though this maybe a nice to have)

Data Splitting

Split the data into a training set and validation set with an 80-20 split.

Model training and evaluation

I will start with the simple model architecture first before training and evaluating it. Then iterate this process trying different architectures and hyper-parameters to reach an accuracy score we are happy with.

8 References

[1] Justin Salamon, Christopher Jacoby and Juan Pablo Bello, “Urban Sound Datasets”, “UrbanSound8K” https://urbansounddataset.weebly.com/urbansound8k.html

[2] Mel-frequency cepstrum Wikipedia page
https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

[3] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research”
http://www.justinsalamon.com/uploads/4/3/9/4/4394963/salamon_urbansound_acmmm14.pdf