Machine Learning Engineer Nanodegree

Urban Sounds Classification using Deep Learning

Valiveti Aditya Pramith Krishna

August 20th, 2019

Domain Background

Sound is all around us. Either directly or indirectly, people are always in contact with sounds. Sounds outline the context of our daily activities, ranging from the sound of a breath when we are alone to the music we dance to when we are in a party, and the other environmental sounds that we hear in every day such as a dog barking, the sound of the rain and the sound of traffic for an urban person. The human brain continuously processes and understands this sound subconsciously, giving us information about the environment around us without much of a thought.

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, but the work on the classification of environmental sounds is comparatively scarce. The problem here is applicability of the Neural Network techniques in the domains, such as sound classification, as many of the discrete sounds happen all the time.

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

There are a large number of real-world applications where audio classification can be used such as:

- Noise-cancelling while audio recording to remove any background noises.
- Security Purposes where we can call emergency services in case of a disaster based on the sounds.
- Automotive where recognizing sounds both inside and outside of the vehicle can improve the safety and comfort of the travelers.
- Identifying different kinds of sounds to assist people with disabilities.
- Content-based multimedia indexing and retrieval

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 which in my case is WAV 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 was detected, we will be presented with an unknown score.

Datasets and Inputs

For this project, we will use a dataset called Urbansound8K [1]. The dataset contains 8732 sound extracts (<=4s) of urban sounds from 10 classes, which are as follows air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, street_music.

The metadata accompanying the dataset contains a unique ID for each sound extract along with its given class name.

These sound extracts are digital audio files in .wav format. Sound waves are digitized by sampling them at discrete intervals known as the sampling rate of 44,100 times per second or 44.1 kHz typically. 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 where typically 16bit which means a sample can range from 65,536 amplitude values. Therefore, the data we will be analyzing for each sound extracts is essentially a one dimensional array or vector of amplitude values.

The data set is unbalanced as there are 1000 samples for all the classes and for siren we have 929 which not a huge difference when considered in a data set of 8732 samples but for car_horn and gun_shot classes there are considerably less samples at 429 and 347.

Solution Statement

The proposed solution is to apply Deep Learning techniques that have proved very successful in the field of image classification in sound or audio classification.

First, we will extract Mel-Frequency Cepstral Coefficients (MFCC) [2] from the audio samples on a per-frame basis with a window size of a few milliseconds. MFCCs are coefficients that collectively make up Mel-frequency Cepstrum (MFC).MFCC summarizes the frequency distribution across the window size, so it is possible to analyze 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.

Benchmark Model

For the benchmark models, 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 an audio slice maximum at duration of 4 seconds.

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

Evaluation Metrics

The evaluation metric for this problem is simply the Accuracy Score. Even though the data is unbalanced it should be fine as we will probably not be fine tuning our model and it I believe that remaining classes will good accuracy except for the ones which have less samples.

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 all audios to have the same sample rate and bit depth.
- Make sure the sample length is uniform (it is already uniform as all the samples are only 4 sec long).

Data Visualizing:
I am thinking of using Librosa for converting audio into arrays for visualizing the audio patterns as wave forms using Matplotlib.

Data Splitting:

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

Model training and evaluation:

For feactures we will use MFCC as mentions in the solution statement. We start with the simple model architecture then train and evaluate it. Then iterate this process trying different architectures and hyper-parameters to reach an accuracy score we are happy with.

The models I am thinking of using are Perceptron's and CNN's to train.

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

[4] S. Chaudhuri and B. Raj. Unsupervised hierarchical structure induction for deeper semantic analysis of audio. In IEEE ICASSP, 2013.