Udacity MLND Capstone Project Proposal

Deeparth Gupta,
December 23, 2019

Classifying urban sounds using deep learning models.

Domain Background

An average person's experience of life is filled with sound. Whether engaged in a conversation or going for a walk, out ears are always picking up sound and our brains are always processing them. It is an important adaptation too. Sound carries a lot of information about the environment that cannot be gleaned with just vision. Allowing any creature capable of hearing to be aware of their surroundings beyond their line of sight.

Automatic environmental sound classification is a growing area of research with many real-world applications. It is still nascent compared to other areas related to speech and music, hence, literature on it is relatively scarce. It could be useful to use techniques used in other domains in this one. In fact, there are examples of this happening.

One such example is the usage of Convolutional Neural Networks in sound processing and classification. Due to their ability to glean spatial relationships between features in an image, they are a great tool to analyze images and the spatial relationship of elements in them. One can train a CNN to analyze spectrographs of various sounds and classify them. CNNs are also able to classify sounds at very high accuracies even if there are multiple sounds mixed together.

Problem Statement

The main objective of this project is to classify common urban sounds using a deep learning model.

When input in the form of a short duration audio file is received, the project should be able to classify it into one of the target sounds.

Datasets and Inputs

The dataset used is called the UrbanSound8K dataset. This dataset contains 8732 sound clips(<=4s) belonging to 10 classes, namely:

- Air Conditioner
- Car horn
- Children playing
- Dog bark

- Drilling
- Engine idling
- Gunshot
- Jackhammer
- Siren
- Street music

The dataset also contains metadata with a unique ID for each of the classes along with other information not required for this project.

Each audio file is a .wav file. Generally, sound is recorded and sored at a sample rate (the rate at which the amplitude of sound is recorded) of 44.1KHz with a bit-depth of 16(the number of bits used to represent each sample). Therefore, a sound signal is essentially a 2D array of amplitude and time.

Solution statement

The idea is to use a Deep CNN to classify sounds as belonging to one of the 10 classes. One way of doing it to represent audio in the form of a spectrogram which is essentially an image. The spectrograph can then be input to a CNN as an image.

One type of spectrogram is called a Mel Spectrogram. It summarizes the frequency distribution across time, so it is possible to analyze both the frequency and time characteristics of the sound.

The next step will be to train a Deep Convolutional Neural Network to recognize representations of these sounds and make predictions about their source. Since they are being supplied with spectrographs, they should perform very well given their capabilities.

Benchmark Model

The algorithms to be compared against are outlined in the original paper for this dataset "*A Dataset and Taxonomy for Urban Sound Research*" by Salamon et al. The paper describes 5 different algorithms with their accuracies when trained on the dataset.

Algorithm	**Accuracy**
SVM with RBF kernel	68%
RandomForest500	66%
IBk5	55%
J48	48%
ZeroR	10%

Evaluation Metrics

The main evaluation metric for this problem is Accuracy.

Project Design

Data Preprocessing

The data is likely to be non-uniform in terms of size, length, sampling rates and bit-depth and will be unsuitable for direct input into a neural network. It will be preprocessed in the following ways.

- Resample all audio to the same sample rate and bit depth.
- Resize all audio samples to the same size
- Consider augmentation such as noise addition and pitch shift.
- Split the dataset into training and validation sets with a ratio of 80/20

Model training and evaluation

Model training should be straightforward. The architecture will be refined before submission.

The model will be trained using cloud-based compute. Google Colab is the most likely to be used. Kaggle and my own desktop are also options.

References

1. Justin Salamon, Christopher Jacoby and Juan Pablo Bello, "Urban Sound Datasets", "UrbanSound8K"
2. Mel-spectogram.
3. J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research"
4. Urban Sound Classification using Convolutional Neural Networks with Keras: Theory and Implementation