Machine Learning Engineer Nanodegree Capstone Proposal

Environmental Sound Classification

Akshay Bhardwaj

November 2nd, 2018

Domain Background

The problem of environmental sound classification is a prominent area of research with it having applications in a platitude of areas, ranging from context aware computing and surveillance to noise mitigation enabled by smart acoustic sensor networks. It is even utilised in content-based multimedia indexing and retrieval.

A variety of signal processing and machine learning techniques have been applied to the problem, including matrix factorization [1], dictionary learning [2], wavelet filterbanks [3] and most recently deep neural networks [4]. The most recent techniques have involved using deep Convolutional neural networks (CNN) [5], and have been particularly successful as they are capable of capturing energy modulation patterns across time and frequency when applied to spectrogram-like inputs, which has been shown to be an important trait for distinguishing between different, often noise-like, sounds such as engines and jackhammers [6].

I was intrigued by this problem after learning about the work done by Google Brain researcher Sara Hooker in this field. She worked with non-profit organizations in Kenya, utilizing old mobile phones to identify chainsaw noises in Kenyan rainforest and thus helping reduce illegal deforestation.

I will be using the audio dataset URBANSOUND8K for this project, which can be found here:
https://urbansounddataset.weebly.com/urbansound8k.html

Problem Statement

The goal of this project is to evaluate the performance of deep-learning CNNs for the classification of urban sonic events. We will try to utilize the power of image classification of CNNs to do this by using the novel approach of turning audio files to spectrogram images and then using transfer learning to classify them. This is done because we want to utilize the recent advancements made in image classification and use transfer learning. We will be using 10-fold cross validation using the predefined folds provided in the dataset to measure the performance of the classifier.

Datasets and Inputs

We will be using the URBANSOUND8K dataset for this project, which can be found here: https://urbansounddataset.weebly.com/urbansound8k.html

This dataset contains 8732 labelled sound excerpts (<=4s) of urban sounds from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music. The classes are drawn from the urban sound taxonomy.

In addition to the sound excerpts, a CSV file 'UrbanSound8k.csv' containing metadata about each excerpt is also provided. This file contains meta-data information about every audio file in the dataset.

The meta-data contains 8 columns.

- slice_file_name: name of the audio file
- fsID: FreesoundID of the recording where the excerpt is taken from
- start: start time of the slice
- end: end time of the slice
- salience: salience rating of the sound. 1 = foreground, 2 = background
- fold: The fold number (1–10) to which this file has been allocated
- classID:
  0 = air_conditioner
  1 = car_horn
  2 = children_playing
  3 = dog_bark
  4 = drilling
  5 = engine_idling
  6 = gun_shot
  7 = jackhammer
  8 = siren
  9 = street_music
- class: The class name: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, street_music.

All excerpts are taken from field recordings uploaded to www.freesound.org. Every recording was manually checked by listening to it and inspecting the user-provided metadata. Only those recordings were kept that were actual field recordings where the sound class of interest was present somewhere in the recording. Next, the audio clips were sliced into short audio snippets for sound source identification of the desired classes. They were then allocated to 10 different folds. Some of the excerpts are from the same original file but different slice. If one slice from a certain recording was in training data, and a different slice from the same recording was in test data, this might increase the accuracy of a final model falsely. Thanks to the original research for which the dataset was created, this has also been taken care of

by allocating slices into folds such that all slices originating from the same Freesound recording go into the same fold. So there is no need to split the data again into training/validation/testing sets.

To avoid large scale differences in the class distribution, the dataset has a limit of 1000 slices per class, with only siren (about 900), car_horn (about 500) and gun_shot (about 400) sounds having lesser number of samples.

Solution Statement

For the solution, we will be trying the novel approach of converting the audio files to their respective spectrogram images and then using image classification on them. This is done because I want to leverage the recent advancements made in image classification to classify these sound files and see if we can get comparable or better performance as was observed when using the audio files directly. I plan to use the 'resnet' architectures for this task instead of the latest inception models as they usually have comparable performances but the resnet models tend to be quite faster.

We will be using the predefined 10 folds provided in the dataset and perform 10-fold cross-validation to evaluate the model as described by the author of the dataset as it is a reliable metric to compare previous results.

Benchmark Model

To have an indication of how our hyper-parameter tuning affect the performance of our model we will use a vanilla 'resnet' CNN architecture (no tuning) as a **base benchmark** model.

Moreover, as it is a novel approach to this problem, we will be using the results obtained by previous researchers using different techniques as a comparative model. We will be trying to match or better the state-of-the-art accuracy achieved by the latest publication (https://arxiv.org/abs/1608.04363) on the dataset website.

Evaluation Metrics

Since the dataset is quite balanced, we will use accuracy as our evaluation metric. The classification model will be evaluated via *10-fold cross validation using predefined splits*. This will enable us to compare our performance to those performed by the author of the dataset as well as later experiments.

**10-fold cross validation using the predefined folds:* train on data from 9 of the 10 predefined folds and test on data from the remaining fold. Repeat this process 10 times (each time using a different set of 9 out of the 10 folds for training and the remaining fold for testing). Finally report the average classification accuracy over all 10 experiments (as an average score, or as a boxplot).

Project Design

The project will be composed of the following steps:

- **Data Download and organisation into folders:** The data organisation will be intermingled throughout the project and will be done as and when necessary.
- **Data preprocessing**: Since all audio samples are 4 seconds long, the dataset fits well with our spectrogram method. I will use the librosa library to convert the audio files to spectrograms. I plan to create a mel-scaled spectrogram, using the librosa.feature.melspectogram function to do so. I do not plan to do any data augmentation as I am unsure about its effect on the spectrograms as unlike normal images, any stretch, zoom, flip, rotate, etc. transformation may correspond to a different category of sound waves (not necessarily in the dataset) and may lead to a worse model.
- **Data Visualization**: Will visualize the converted images of different audio types to see if any pattern is observable to the human eye.
- **Model training and evaluation**: I plan to use transfer learning, and use the pre-trained resnet34 architecture for initial model training. But since the pre-trained model is accustomed to the Imagenet dataset, I will later unfreeze all the layers of the model and then train the whole architecture with a differential learning rate. I intend to use the fastai library to do so as it allows for easy flexibility in training the CNN layers with differential learning rates. The models will be trained and evaluated on different folds separately to get their individual performances and then they will be averaged out to get the 10-fold cross validation accuracy.
- **Conclusion**

References

[1] - V. Bisot, R. Serizel, S. Essid, and G. Richard, “Acoustic scene classification with matrix factorization for unsupervised feature learning,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 6445–6449

[2] - J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 171–175.

[3] - J. T. Geiger and K. Helwani, “Improving event detection for audio surveillance using gabor filterbank features,” in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, Aug. 2015, pp. 714– 718

[4] - K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, Sep. 2015, pp. 1–6.

[5] - J. Salamon and J. P. Bello, “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification” in IEEE Signal Processing Letters, November 2016

[6] - ——, “Feature learning with deep scattering for urban sound analysis,” in 2015 European Signal Processing Conference, Nice, France, Aug. 2015