"Alexandru Ioan Cuza" University of Iaşi

Master of computational optimization

FACULTY OF COMPUTER SCIENCE

Advanced Software Engineering Techniques 2019 Project

- State of the art -

Freesound Audio Tagging 2019

Automatically recognize sounds and apply tags of varying natures

Proposed by: Cojocaru Gabriel-Codrin
Dinu Sergiu Andrei
Luncaşu Bogdan Cristian
Racoviţă Mădălina-Alina
Vîntur Cristian

Scientific coordinators: PhD Associate Professor Adrian Iftene
PhD Associate Professor Mihaela Elena Breaban

Contents

Contents **2**

State of the art **3**

Problem description 3

Short audio files classification 3

Problem statement 3

Dataset description 3

Proposed solution 4

Results 5

Tagging longer audio files 5

State of the art 5

Relevant articles 6

Relevant articles 6

Bibliography **6**

State of the art

Problem description

The environmental sound classification problem can come in many different shapes, from having to classify short audio files with a label from a specified set to having to tag live audio streaming with one or multiple labels.

Automatic environmental sound classification or tagging is a growing area of research. Work done in this area is comparatively scarce with work done in related audio fields such as speech and music. Applications are numerous and include multimedia indexing and retrieval, environmental sounds subtitles, assisting deaf individuals or even monitoring illegal deforestation. A very interesting project regarding this last use case is created by a non-profit organization called Delta Analytics. They helped build a system where a lot of old mobile phones are attached to trees in rain forests which listen to chainsaw noises. Their role is to identify when a chainsaw is being used and alert rangers who can stop illegal deforestation.

Because of the recent success in image classification, the question that raises is: can we bring techniques used in image classification to sound classification by representing sound in various image formats?

Short audio files classification

Problem statement

Classify environmental sounds with focus on identification of particular urban sounds. Given an audio sample of a few seconds duration determine if it contains one of the target urban sounds.

Dataset description

The dataset is called Urbansound8K and contains 8732 sound excerpts (under 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

By taking a closer look to the dataset, the following observation can be made: it's tricky to visualize the difference between some sounds, especially the continuous ones like jackhammer and engine idling.

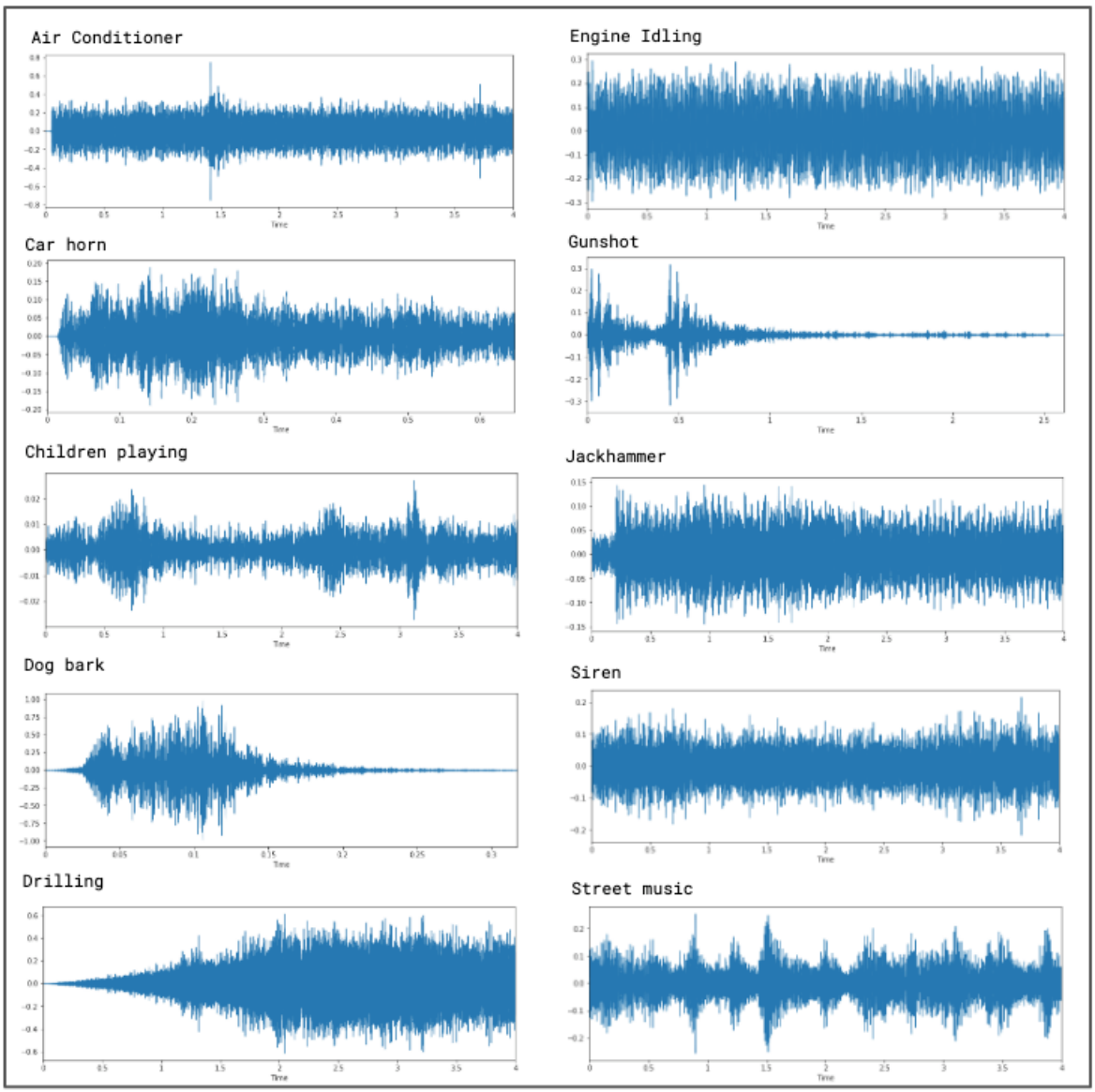

Fig. : *Visualization of sound amplitude with respect to time*

A deeper dive into the data show that: most samples have 2 audio channels with a few having just 1 channel, sample rate varies from 8kHz to 48kHZ and bit depth also varies from 4bit to 32bit which leads to requiring a data normalization step.

Proposed solution

The proposed solution involves mapping the audio files, after the normalization step, to a visual representation. Spectrograms are a useful technique for visualizing the spectrum of frequencies of a sound and how they vary during a very short period of time. A similar technique known as mel-spectrogram will be used for this case. The main difference is that a spectrogram uses a linear spaced frequency scale (so each frequency bin is spaced an equal number of Hertz apart), whereas a mel-spectrogram uses a quasi-logarithmic spaced frequency scale, which is more similar to how the human auditory system processes sounds.

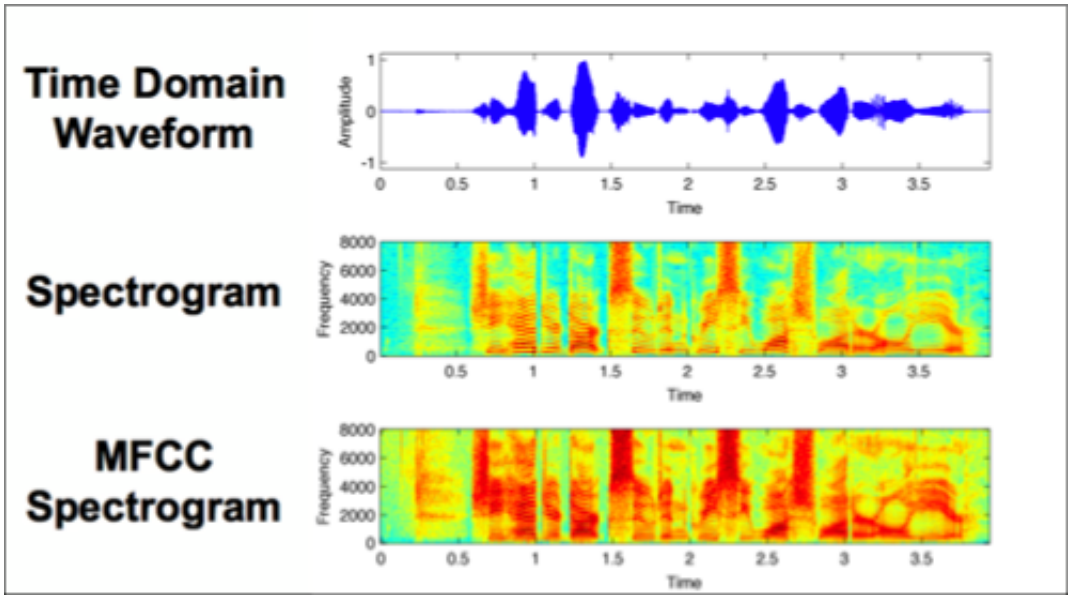

Fig. : *Spectrogram and mel-spectrogram*

The problem will be solved by using deep learning on the visual representation of the audio file. In this case a Convolutional Neural Network (CNN) will be used. The model used is a sequantial one consisting of 4 Conv2D convolution layers with the final output being a dense layer. The output layer will have 10 nodes which matches the number of possible classifications.

Results

The trained model obtained a Training accuracy of 98.19% and a Testing accuracy of 91.92%.

Tagging longer audio files

By providing a solution to solving the short audio files tagging problem we can extend it to work on longer audio files. The idea behind this extension is to divide the audio into smaller parts, transform these parts into mel-spectrograms and feed them into a more complex network that will be able to provide the correct label(s).

State of the art solution

State of the art solution to solving this problem is deep learning. By dividing the audio and transforming it into mel-spectrograms we can use some of the success of image classification

and bring it to audio classification. Just applying image classification on mel-spectrograms doesn't work because sound in form of mel-spectrograms doesn't have the same properties as an image but by adapting the network the the new context we can obtain a solution to solving this problem.

Relevant articles

- AUDIO TAGGING WITH NOISY LABELS AND MINIMAL SUPERVISION - *https://arxiv.org/pdf/1906.02975.pdf*
- Convolutional RNN: an Enhanced Model for Extracting Features from Sequential Data - *https://arxiv.org/pdf/1602.05875v3.pdf*

Relevant links

- Kaggle competition 1st place solution - *https://github.com/lRomul/argus-freesound*
- Kaggle competition 2nd place solution - *https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97815*
- Kaggle competition 3rd place solution - *https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97926*

Bibliografie

- **"Classifying Urban sounds using Deep Learning"**, Mike Smales: *https://github.com/mikesmales/Udacity-ML-Capstone/blob/master/Report/Report.pdf*
- **"Data Science for Good: Stopping Illegal Deforestat"**, Sara Hooker, Sean McPherson: *https://mlconf.com/sessions/data-science-for-good-stopping-illegal-deforestat-2/*
- Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Serra. **"AUDIO TAGGING WITH NOISY LABELS AND MINIMAL SUPERVISION"** in Detection and Classification of Acoustic Scenes and Events 2019