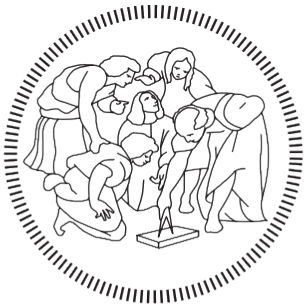
Academic Year 2022/2023



POLITECNICO
MILANO 1863

Politecnico di Milano

Numerical Analysis for Machine Learning

Urban Sounds Classification

Igor Katselenbogen - 10855055
Leonardo La Rocca - 10930221

Prof. Edie Miglio

January 15, 2023

Contents

1 Introduction

In this project, we set out to create a machine learning model that could accurately classify a range of different environmental sounds. To achieve this goal, we employed the use of urban sound excerpts from the UrbanSound8K dataset and implemented a convolutional neural network model using the Keras library. In an effort to improve the performance of our model, we also investigated the application of four audio data augmentation techniques. This document presents the details of our implementation and the results of our experiments.

2 Technologies

- **Python** — a general-purpose programming language often used in scientific computing and data analysis.
- **Librosa** — a Python library for audio and music processing.
- **Muda** — a Python library for musical data augmentation.
- **Keras** — is an open-source machine learning library for Python.
- **Numpy, Pandas, Matplotlib, Seaborn** — are popular Python libraries for data analysis and visualization.

3 Project Description

The project consists of two main parts:

- **Data preprocessing and augmentation**
- **Model implementation and training**

During the data preprocessing step, we converted the audio samples from the dataset into log-mel spectrograms. Spectrograms are a visual representation of sounds that capture the temporal and spectral characteristics of sound signals. Spectograms can be used as inputs to train 2D convolutional neural networks that require the data to be in an image-like shape.

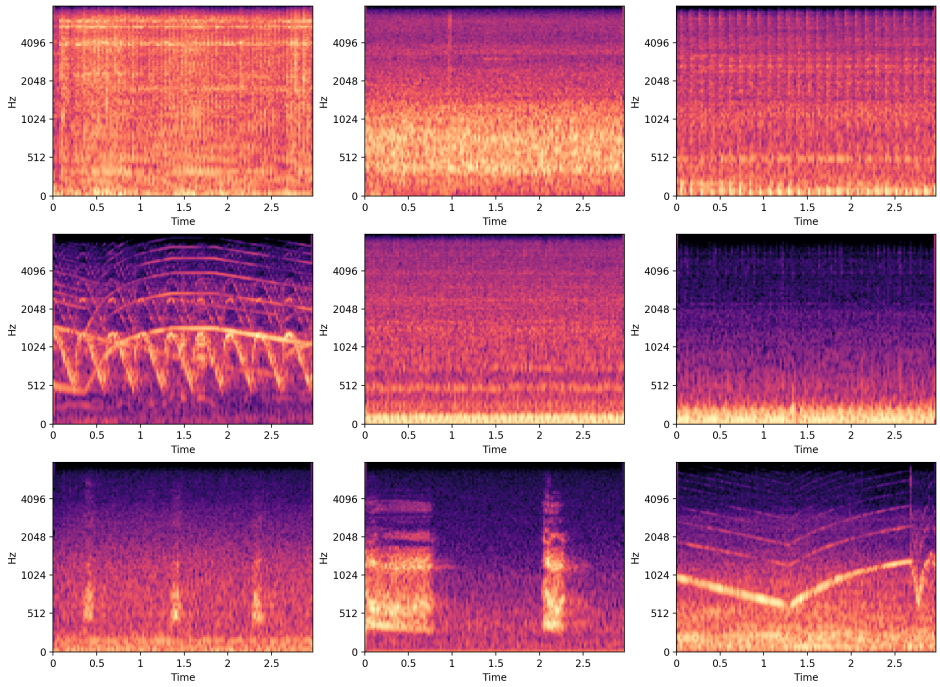


Figure 1: Examples of log-mel spectograms from the dataset

To expand the size of our training data, we applied four data augmentation techniques, that were specifically designed for audio data. We augmented the data before transforming it into images.

For the model, we implemented a simple convolutional neural network (CNN) architecture, comprising three convolutional layers for the feature extraction part and two dense layers for the classification part.

We evaluated the model using a 10-fold cross-validation method on predefined and stratified subsets of the dataset. We used loss and accuracy to assess the performances of the model.

4 UrbanSound8K Dataset

The UrbanSound8K dataset has been widely used in research on urban sound classification, particularly for evaluating the performance of machine learning models on this task. It has also been used to explore other aspects of sound analysis, such as feature extraction and sound event detection.

The UrbanSound8K consists of annotated urban sound recordings. It contains 8732 labeled sound excerpts of urban sounds from 10 different classes:

- air_conditioner
- car_horn
- children_playing

- dog_bark
- drilling
- engine_idling
- gun_shot
- jackhammer
- siren
- street_music

Each sound excerpt is an audio clip labeled with the class of the sound it contains. Sound excerpts were originally extracted using environmental field recordings uploaded to www.freesound.org.

The original environmental audio samples were recorded in various locations such as New York City's streets, parks, and public transportation. The sound excerpts in the dataset are one second long and are sampled at 44.1kHz. The dataset is split into 10 folds, with each fold containing 873 sound samples distributed among all the classes, allowing for cross-validation when training machine learning models on the dataset.

The UrbanSound8K dataset includes a metadata CSV file that provides class labels for each sound file as well as other information.

	slice_file_name	fsID	start	end	salience	fold	classID	class
0	100032-3-0-0.wav	100032	0.0	0.317551	1	5	3	dog_bark
1	100263-2-0-117.wav	100263	58.5	62.500000	1	5	2	children_playing
2	100263-2-0-121.wav	100263	60.5	64.500000	1	5	2	children_playing
3	100263-2-0-126.wav	100263	63.0	67.000000	1	5	2	children_playing
4	100263-2-0-137.wav	100263	68.5	72.500000	1	5	2	children_playing

Figure 2: Rows from the metadata file

The metadata file contains both numerical and textual class labels. Each sound file was extracted from a longer recording, and the metadata file includes the start and stop times of the extracted portion.

The metadata also includes a salience feature, which refers to the prominence or noticeability of the sound within the recording. This binary feature is indicated by two states: foreground (1) and background (2).

5 Data augmentation

To compensate for the relatively small size of the dataset, four data augmentation techniques were used. Data augmentation was performed offline by applying the techniques on the audio samples before they were converted to spectrograms. All the techniques were applied on each sample to create four augmented datasets.

The four techniques we implemented were:

- **Background Noise** — adds background noise to the audio sample. The background noises consisted of four audio recordings of various urban acoustic scenes, such as street and park noises. Each sample was combined with each background noise recording to generate four new augmented samples.
- **Time Stretching** — changes the speed of the audio sample but keeps the pitch unchanged. Each sample has been time-stretched by the following factors: 0.81, 0.93, 1.07, 1.23.
- **Pitch Shifting** — shifts the pitch of the audio sample by several semitones, changing the tonality of the sounds. Each sample has been pitch-shifted by the following factors: 3.5, 2.5, 2, 1, 1, 2, 2.5, 3.5.
- **Dynamic Range Compression** — an operation that reduces the volume of loud sounds and amplifies quiet sounds, thus reducing or compressing an audio signal's dynamic range. Each sample has been processed with four different sets of parameters for the dynamic range compression: music standard, film standard, speech, radio.

We have augmented each sample using all the techniques. As a result, the augmented dataset consists of 183372 samples.

6 Model

For the model, we implemented a Convolutional Neural Network architecture with the following layers and parameters:

- Layer 1: 24 filters with a receptive field of (5,5). This is followed by (4,2) strided max-pooling over the last two dimensions (time and frequency respectively) and a ReLU activation function.
- Layer 2: 48 filters with a receptive field of (5,5). Like layer-1, this is followed by (4,2) strided max-pooling and a ReLU activation function.
- Layer 3: 48 filters with a receptive field of (5,5). This is followed by a ReLU activation function (no pooling).
- Layer 4: 64 hidden units, followed by a ReLU activation function and dropout layer.
- Layer 5: 10 output units, followed by a softmax activation function.

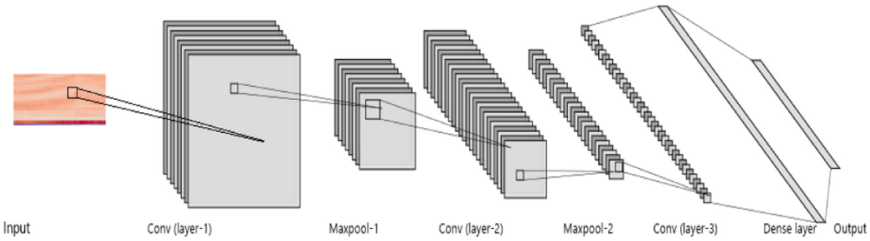


Figure 3: The architecture of the model

7 Results

To measure the model's effectiveness, we used a 10-fold cross-validation method. This is necessary because samples from the same class within a folder are from the same source audio files. This could lead to inflated results if related samples are placed in both the train and test sets. To avoid this, we used different folds for training and validation.

The evaluation metrics we used were loss and accuracy. The loss function was categorical cross-entropy loss, and the accuracy was calculated as the percentage of correctly classified instances.

We evaluated the accuracy of the model both on the original dataset and the augmented dataset. The following table illustrates the results that we observed:

Validation Fold	Accuracy No Aug.	Loss No Aug.	Accuracy Aug.	Loss Aug.
Fold 1	0.6057	1.3375	0.7081	1.4160
Fold 2	0.6478	1.1808	0.7094	0.9995
Fold 3	0.6350	1.2900	0.6699	1.1197
Fold 4	0.6468	1.2933	0.7183	1.0366
Fold 5	0.7321	0.9703	0.7934	0.7471
Fold 6	0.6796	1.1647	0.7332	1.0983
Fold 7	0.6262	1.3124	0.7038	1.0634
Fold 8	0.6497	1.2885	0.6595	1.2144
Fold 9	0.7552	0.9662	0.7672	0.9483
Fold 10	0.7500	1.0129	0.8111	0.7251

It's clear from the observed data that the assessment metrics improved when we augmented the dataset. More specifically, we have obtained improvements on every fold, except for the first one with regards to the Loss metric, which got worse by 5.5%. On average, the accuracy improved by 5.6%, and on average reached 72.1% .
To conclude, we can confirm that applying augmentation techniques is essential when working with audio information. In particular, the proposed data augmentation techniques proved to be effective.

References

[1] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification", submitted, 2016