Assignment - 1 : Audio Classification

Aditya Jain
Department of Electrical Engineering,
Indian Intitute Of Technology Kanpur
adityajain20@iitk.ac.in

1 Introduction

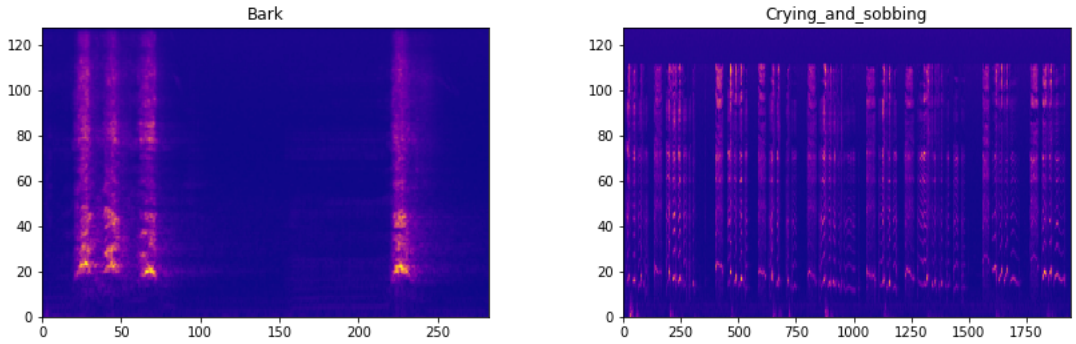
In recent years the classification of sound or audio recognition system has expanded its momentum in various fields like different animals voice recognition, automatic screams detection, the combination of video and audio for crime scene warning system, IoT based solution for urban noise detection in smart cities, sound classification with detection for medical and health care problems, classification of distinct musical instruments and many more. This shows the importance and scope of autonomous sound recognition systems in almost every aspect of not only humans but also other's living organisms like trees and animal's life [1]. This assignment is particularly based on classification of ten distinct audio - Crying and sobbing , Vehicle horn and car horn and honking, Doorbell, Knock, Siren, Bark, Walk and footsteps, Shatter, Meow and Microwave oven, on the basis of their mel-spectrogram.

2 Literature Survey

In past, there have been numerous attempts for environment sound classification (ESC) using various models, inclusive of , but not limited to Gaussian Mixture Models (GMMs), Dense Neural Network (DNNs), Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Support Vector Machines and K-Nearest Neighbours algorithm based approaches. First works related to CNN on ESC were conducted by Piczak. Piczak designed a CNN architecture which has two convolutional layers, two max pooling and two fully connected (FC) layers. In addition, Piczak constituted ESC-10 and ESC-50 databases, and these databases were used to evaluate classification performance of the CNN architecture. The reported accuracy scores for ESC-10, ESC-50 and UrbanSound8k were 81.0%, 64.5% and 74%, respectively. Salamon et al. improved the classification performance by boosting of the convolutional layers and using data augmentation, which includes time stretching, pitch shifting, dynamic range compression and background noise. The classification accuracy with this method using UrbanSound8k database was 74% without data augmentation. Aytar et al. proposed the SoundNet, which contains the 1D CNN architecture. Chen et al. used an approach called dilated convolutions for ESC. In the proposed method, dilated convolutions were preferred instead of max-pooling. Authors used UrbanSound8k database for performance evaluation of the proposed method. The obtained accuracy score was 78.0% [2].

3 Methods

The primary step towards building an efficient classification model was to visualize and clean all the available data.



As observable from the above spectrograms, there is a lot of empty space in the data and also there is a repetition of pattern.

3.1 Pre-processing

Various pattern recognition techniques have been employed to process the available data.

3.1.1 Method-1

Extracted K number of columns, in descending order, on the basis of the sum of the values present in the column and form a new matrix of size 128 x K , while conserving the order in which column were present in the original matrix.

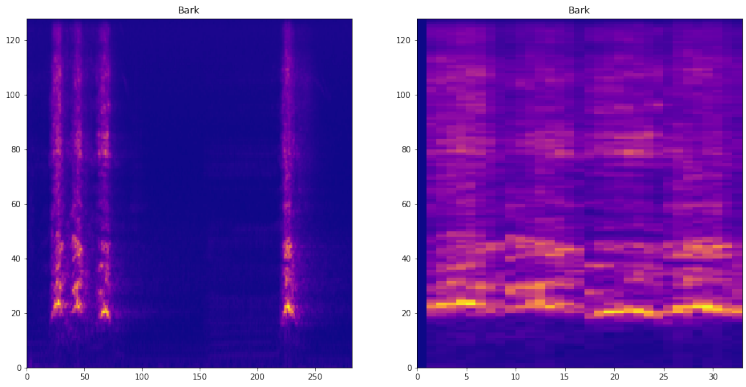
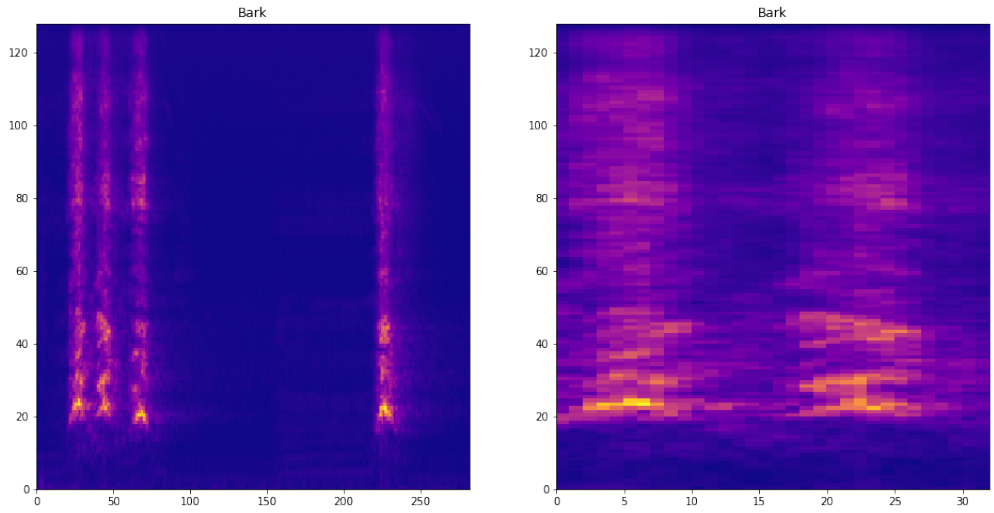


Figure 1: Original(left) and processed(right) spectrogram

3.1.2 Method-2

Of all the repeated patterns, take only 1 pattern and form a new matrix of size 128 x K .This is implemented by choosing submatrix of dimension 128 X K which has maximum cross-correaltion with the original matrix.



Although, this process is very accurate, it is computationally extensive. To reduce the computations, a little less accurate but 100 times faster method was used.

3.1.3 Method-3

This method is a faster implementation of method-2. Instead of finding the repeated pattern by cross correlating 2-D matrices, the pattern was found by cross correlating 1-D arrays containing the sum of values of each column of the 2-D matrix.

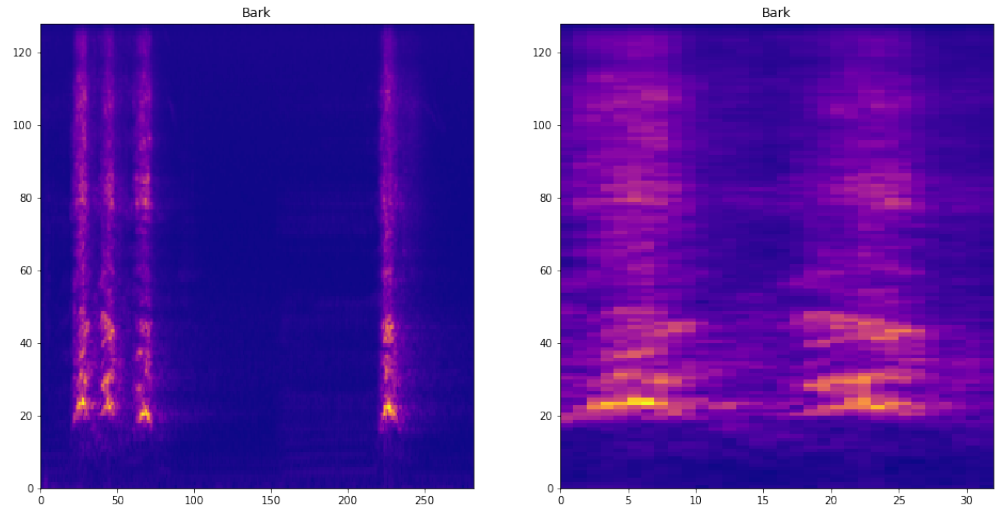


Figure 2: Original(left) and processed(right) spectrogram

3.2 Data Augmentations

The available data was not sufficient for training an efficient CNN model. So, blur and line augmentation was applied to the processed spectrograms to generate a dataset with 200 samples of each class.

3.3 Models

Two different approaches were used to build the classifier.

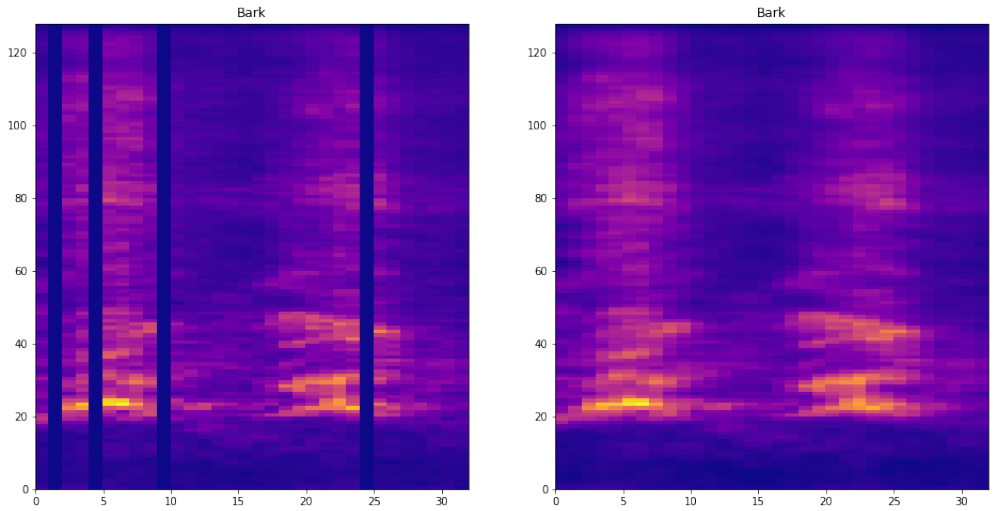


Figure 3: line augmented (left) and blurred(right) spectrograms

3.3.1 Dense Neural Network

A fully connected dense neural network with 6 dense layers and multiple dropout layers with dropout probability of 0.2 has been created. Further details of the architecture are given in Fig 4.

Layer (type)	Output Shape	Param #
flatten_2 (Flatten)	(None, 4096)	0
dense_4 (Dense)	(None, 4096)	16781312
dropout (Dropout)	(None, 4096)	0
dense_5 (Dense)	(None, 1024)	4195328
dropout_1 (Dropout)	(None, 1024)	0
dense_6 (Dense)	(None, 512)	524800
dense_7 (Dense)	(None, 128)	65664
dense_8 (Dense)	(None, 64)	8256
dense_9 (Dense)	(None, 10)	650

Total params: 21,576,010
Trainable params: 21,576,010
Non-trainable params: 0

Figure 4: DNN Architecture

3.3.2 Convolution Neural Network

Since the spectrograms are in the form of 2-D matrices, multilayer perceptrons based CNNs are one of the most viable approach. A CNN with four 2-D convolution layers, three 2-D max-pooling layers and 3 dense layers has been deployed. Further details of the architecture are are given in Fig 5.

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_4 (Conv2D)            (None, 64, 16, 16)        160

max_pooling2d_3 (MaxPooling  (None, 32, 8, 16)         0
2D)

conv2d_5 (Conv2D)            (None, 32, 8, 32)         4640

max_pooling2d_4 (MaxPooling  (None, 16, 4, 32)         0
2D)

conv2d_6 (Conv2D)            (None, 8, 2, 64)          18496

conv2d_7 (Conv2D)            (None, 8, 2, 64)          36928

max_pooling2d_5 (MaxPooling  (None, 4, 1, 64)          0
2D)

flatten_1 (Flatten)          (None, 256)               0

dense_2 (Dense)              (None, 128)               32896

dense_3 (Dense)              (None, 10)                1290

=================================================================
Total params: 94,410
Trainable params: 94,410
Non-trainable params: 0
_________________________________________________________________
```

Figure 5: CNN Architecture

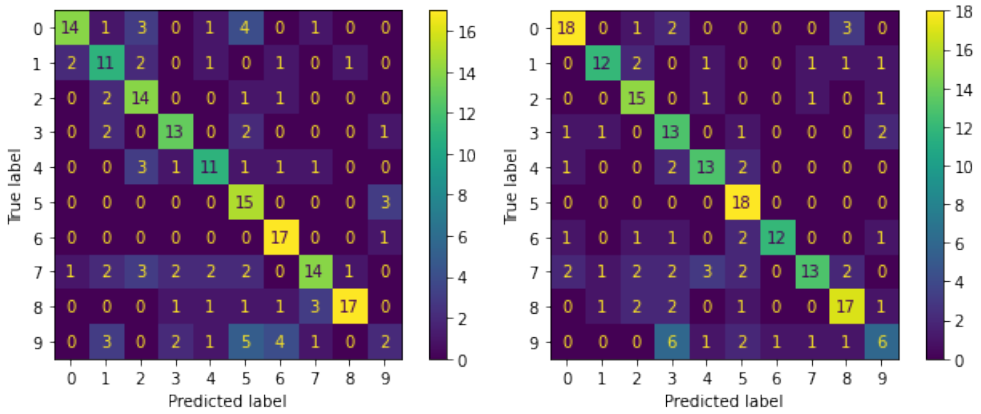
4 Results and Discussion

4.1 Validation Results

	Accuarcy	Precision	Recall	F1_score
DNN Model	0.7852	0.8735	0.7075	0.78
CNN Model	0.9053	0.8768	0.8454	0.8606

4.2 Test Results

	Accuarcy	Precision	Recall	F1_score
DNN Model	0.6368	0.6283	0.6421	0.6178
CNN Model	0.6816	0.697	0.6884	0.6762



Confusion Matrix for DNN(left) and CNN(right)

As observed, the CNN model performs much better than the DNN model due to it translation invarience property. Further, performance of the models could be increased by training it on more data and with better pre-processing techniques.

5 References

[1] ZohaibMushtaq, Shun-FengSu, Quoc-VietTran *Spectral images based environmental sound classification using CNN with meaningful data augmentation.*

[2] Fatih Demir, Muammer Turkoglu, Muzaffer Aslan, Abdulkadir Sengur *A new pyramidal concatenated CNN approach for environmental sound classification*