EasyChair Preprint

№ 8056

Acoustic Scene Analysis and Classification Using Densenet Convolutional Neural Network

Samyak Doshi, Tushar Patidar, Shubhankar Gautam and Rajkishor Kumar

May 24, 2022

ACOUSTIC SCENE ANALYSIS AND CLASSIFICATION USING DENSENET CONVOLUTIONAL NEURAL NETWORK

Samyak Doshi[1], Tushar Patidar[2],Shubhankar Gautam[3] Rajkishor kumar[4]

VIT University, Vellore, India

Samyak.2018@vitstudent.ac.in, tushar.patidar2018@vitstudent.ac.in,
Shubhankar.gautam2018@vitstudent.ac.in, rajkishor.kumar@vit.ac.in

Abstract – Environment sound classification is a type of Sound Event Recognition (SER). Environmental Sound Classification has always been one of the major issues in audio recognition sector. Active researches are going on, in the audio domain and we have seen a lot of progress in the past years. In this paper we present an account of state-of-the-art in Acoustic Scene Classification (ASC), the task of environmental scenario classification through the sounds they produce. Our work aims to classify 50 different outdoor and indoor scenario using environmental sounds. We use a dataset ESC-50 from the IEEE challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). In this we propose to use 2000 different environmental audio recordings. In this method the raw audio data is converted into Mel-spectrogram and then the generated Mel-spectrogram is fed as an input to neural network for training. Our model follows structure of neural network in the form of convolution and pooling. With a focus on real time environmental classification and to overcome the problem of low generalization in the model, the paper introduced white noise into environmental sounds which also act as an audio enhancement.

Keywords – Audio classification, Convolutional Neural Network, Mel-spectrogram, DenseNet

I. INTRODUCTION

Acoustic Scene Classification (ASC) is a progressing field in audio analysis and research. Field. ASC refers to the activity of associating a semantic label to an audio that recognize the environment in which it has been produced. Sound Event Detection is the process of recognizing the sound events with their start and end time in a recording.[1] Sound events do not always occur in isolation in the real life, they tend to overlap each other, especially environmental events. An acoustic scene is made up of several sounds or occurrences. Consider an airport, where there are numerous sound events such as airline announcements, people conversing, phones ringing, children playing, and so on. Individually or in a series of events, the human auditory system is particularly capable of distinguishing between various sounds. Auditory Scene Analysis (ASA) is the term for this [2]. This can be mechanized by teaching robots to recognize distinct real-world events, a process called as Computational Auditory Scene Analysis (CASA) [3]

Distinguishing of environment sounds plays an important role in surveillance and security aspects. ASC can be useful in many upcoming applications such as hearing aids for disabled people, automated analysis of urban sounds, incorporation in autonomous vehicle. It can also be used to predict maintenance for industrial machines or for content - based multimedia indexing and retrieval.

Initially, convolutional neural networks (CNN) were frequently utilized in image categorization. Later, when one-dimensional convolution and sound feature image extraction technologies advanced, CNN progressively became the standard approach of environmental sound classification.[4]

Due to the non-stationary nature of ambient sound and the high level of disturbance, an environmental sound classification algorithm based on adding white noise is proposed in this study to give it a pinch of reality [5]. In acoustic processing, frequency features are highly used. Features such as Mel-spectrogram, Chroma, Tonnetz, LPCC and MFCC are extracted from the audio data. Our paper addresses the use of CNN and DCNN along with a real time audio mixer in the form of augmented data. One of the key applications is detecting unusual activity like as sobbing or shouting in agony in an enclosed space, or shots from a rifle, and so on. This can be accomplished by audio surveillance, which employs sound content analysis tools to find outliers.[6]

II. LITERATURE SURVEY

Various techniques exist for environmental sound classification, Behnaz Bahmei et. al [7] proposed a neural network technique with a composition of RNN and CNN. Additionally the author proposed DCGAN (Deep Convolution Generative Adversarial Network) in order to augment the data.

Wen Zhao et. al [4] introduced a simple neural model with an astonishing accuracy of 94.73% over original audio data without being processed. The paper further introduced an innovative approach of augmentation via SNR noise. Following the implementation in his model, he addressed a gradual decrease of 2% in accuracy in case of low SNR while high SNR increases the generalization of the model.

So to achieve a task with real time analysis, we introduced white gaussian noise with adjustable parameters which covers augmentation of our model.

Wenjie-Mu et. al [8] proposed a temporal-frequency attention based convolutional neural network model(TFCNN) which is a very impressive approach in terms of supervised learning. These mechanisms focuses on key frequency bands

and semantic related time frames on the spectrogram to reduce the influence of background noise and irrelevant frequency bands.

Following the work of Behnaz Bahmei [7], Manjunath Mulimani et. al [9] performed CRNN (Convolution Recurrent Neural Network) as the primary method to classify dataset using 3 activation function Relu, LeakyRelu and ELU. This model overcame the saturation problem over accuracy in a much better way than Behnaz Bahmei [7] model.

III. MODEL OVERVIEW

The model consist of testing the random input audio file for validation followed by preprocessing data and later on to train the model with those datasets and classifying it into different classes. To increase the dataset our paper introduced data augmentation followed by data transformation and graph plotting.

In order to analyze the data further out paper address use of Mel-Spectrogram, Tonnetz, Chroma and MFCC for audio characteristic and for professional usage.

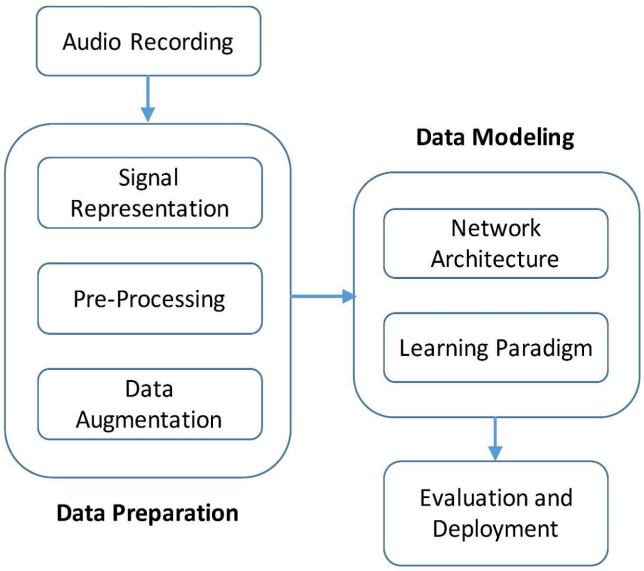

Figure1: Flow Chart for the Proposed Model

Our CNN model consists of 12 layers which are divided into four parts namely pooling layer, filter layer, convolution layer and fully connected layer

A raster scan is run over each pixel in the convolution layer, and pixels are normalized using arithmetic operations. [1] The input image is compared by patching via filters in the filter layers, and negative pixel values are deleted from the matrix. The image size is lowered in the pooling layer by putting the computed matrix through a window filter for feature extraction [5]. Finally, our model includes a fully linked layer in which categorization is accomplished by joining each neuron network in CNN.

IV. PROPOSED METHODOLOGY

The developed approach for ASC is depicted in Fig 2. The steps for demonstrating acoustic scene classification are divided into two parts: feature extraction and audio classification using a convolution neural network.

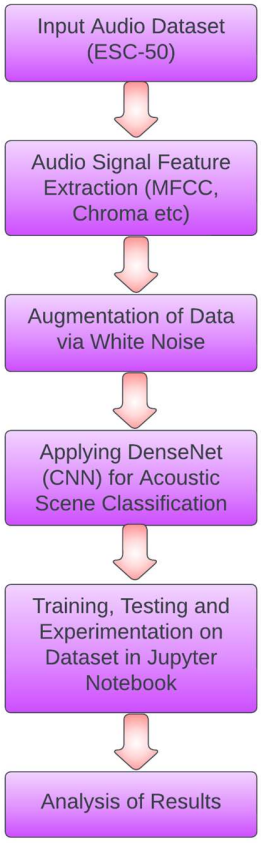

Figure 2: Algorithm of the proposed ASC model

Database description

The ESC-50 dataset is a tagged collection of 2000 environmental audio recordings that can be used to test environmental sound classification methods. The dataset is made up of 5-second-long recordings that are grouped into 50 semantical classes (40 samples per class) that are informally organized into 5 broad categories.

The clips in this dataset were hand-picked from public field recordings acquired by the Freesound.org initiative. The dataset has been pre-folded into 5 folds for similar cross-validation, with fragments from the same original source file stored in a single fold.

Data Augmentation

Data augmentation is a powerful approach for increasing the diversity of available data and allowing models to be trained without the need for additional data. Background noises (crowd sound, church bell, keyboard typing) were added to the data samples (the background noises were acquired from public recordings accessible).

Neural network

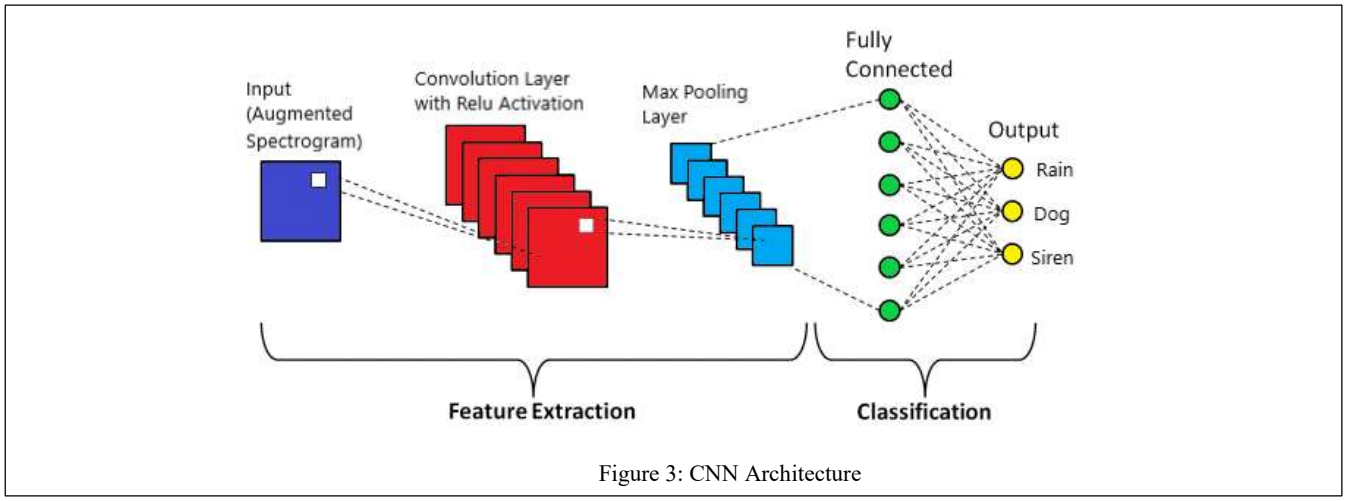

Figure 3: CNN Architecture

A Convolution Neural Network is a powerful Deep Learning algorithm. Pattern recognition in pictures and audio inputs is accomplished using a convolution neural network [10]. The feed-forward function is more efficiently implemented using a convolution neural network.

In this paper we use DenseNet CNN for classification and training of our model. The neural network consist of 4 convolution layers, 4 pooling layers and 1 fully connected layer. After each convolution layer, activation function ReLU is used to decrease the cost of the model by deciding the activation of neurons.

1. The first stage of its algorithm takes place with the convolution operation in which a linear operation is performed. Convolution is conducted on input data to generate a feature map using a filter or kernel[11].

2. At each region, a matrix multiplication is conducted, and the feature map's sums of results are updated. Then convolution is followed by ReLU activation function. In a convolution neural network or deep learning, the ReLU is the most generally utilized activation function. The disappearing gradient is reduced using ReLU. As the dataset grows, there is a significant change in input and a minor change in output [10][12]. As a result, the derivative gets small and the gradient loss function approaches 0, making training the network difficult.

3. After this the transformed data is fed into max pooling layer which aids in reducing the geographic dimension of the input as well as the number of parameters in the network. Pooling is classified into two types: maximum pooling and average pooling. The term "max pooling" refers to a technique that is commonly employed.

4. Max pooling is followed by fully connected layer which helps in the classification of images into labels by utilizing the findings of the convolution and pooling processes[13].

5. Softmax follows the fully connected layer which are used for multiple classification. Softmax is employed as the activation function in multi-class classification issues when class membership on more than two class labels is required.[12]

Subparts of audio analysis – chroma, tonnetz, Mel-spectrogram and MFCC

- Chroma –
  The Chroma Feature is a descriptor that refers to the tonal content of a melodic sound stream in a consolidated form. As a result, chroma highlight might be regarded as a significant requirement for advanced level semantic testing, akin to harmony recognition or consonant closeness estimation [4]. A higher quality of the extracted chroma include enables significantly better outcomes in these higher level assignments. The chroma is calculated by taking into account the log-repeat size range across octaves. The chroma-gram is the resulting plan of chroma vectors.

  Cf(b)=Z1z=0|Xlf(b+z)||

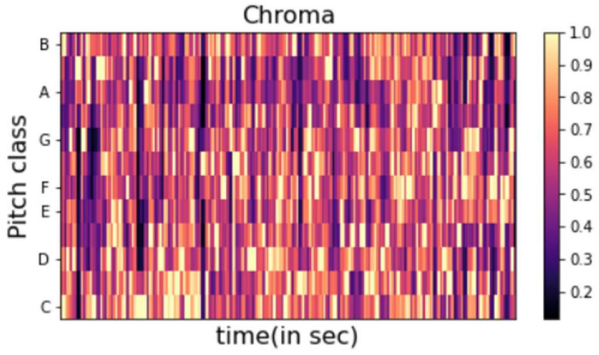

Figure 4: Chroma spectrum of a random audio input

- **Tonnetz –**
  The Tonnetz is a pitch space defined by a network of melodic contribute just inflection links. On a large Euclidean plane, close symphony relationships are shown as short separations.

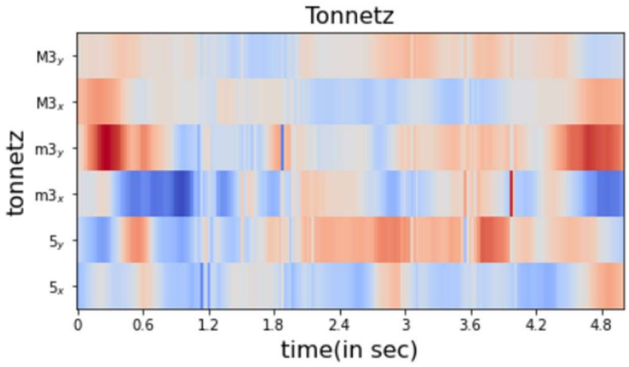

Figure 5: Tonnetz spectrum if a random audio input

- **Mel-Spectrogram**
  A Mel - scale depiction of a signal's frequency spectrum, where the frequency spectrum of a signal is the frequency range that the signal contains. It adopts a linear spaced frequency scale (Short TIME Fourier Transform (STFT)).

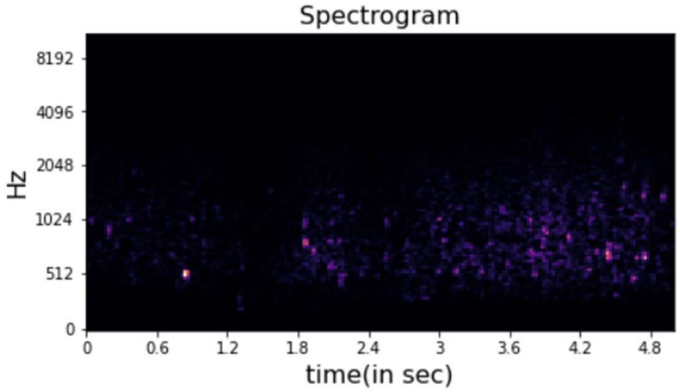

Figure 6: Mel-Spectrogram of a random audio input

- **MFCC**
  MFCC is a popular approach for extracting features from an audio source. MFCC uses a quasi – logarithmic spaced frequency scale. MFCC of an audio set of insects from ESC – 50 datasets has been shown below.

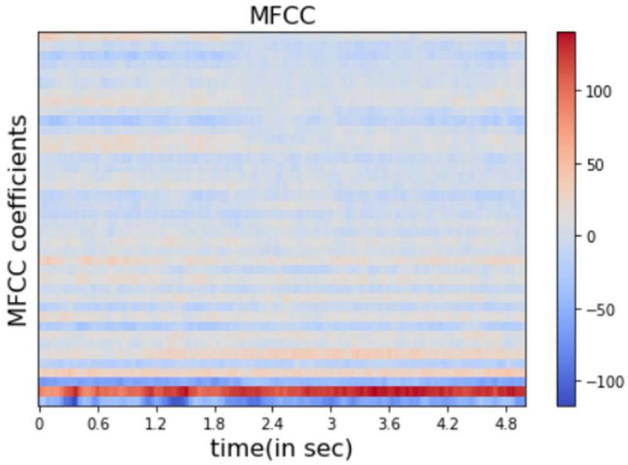

Figure 7: MFCC of a random audio input

V. RESULT AND DISCUSSION

The proposed Acoustic Scene Classification system was subjected to training and testing from ESC-50 dataset which is a part of DCASE 2016. The dataset is made up of 5-second-long recordings that are grouped into 50 semantical classes (40 samples per class) that are informally organized into 5 broad categories. Overall the number of samples is 2000. For testing purpose of the trained DenseNet model, 360 audio samples are chosen.

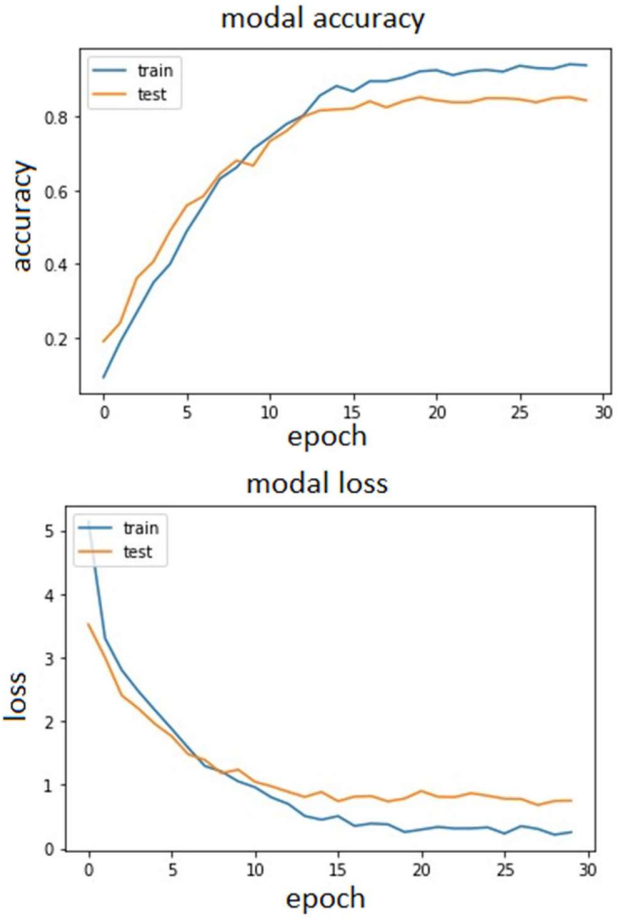

Figure 8: Model accuracy vs epoch graph

The model achieved 86.43% accuracy in validation accuracy and 96.06% in training accuracy.

Corresponding to the result, we can observe that the model is sequential DenseNet CNN and it represents the output shape of array (classification). With 8 hidden layers, each layer being transformed with respect to size and shape along with number of parameters used for training the model can be observed in Figure 9.

Future applications based on this concept could revolve around hearing aid applications for disabled people, integrating features in autonomous cars, contextual based multimedia indexing and retrieval, and predictive maintenance for industrial machinery.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 126, 214, 64)      640
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 63, 107, 64)       0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 61, 105, 128)      73856
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 30, 52, 128)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 28, 50, 256)       295168
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 14, 25, 256)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 12, 23, 256)       590080
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 6, 11, 256)        0
_________________________________________________________________
flatten (Flatten)            (None, 16896)             0
_________________________________________________________________
dense (Dense)                (None, 256)               4325632
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 50)                12850
=================================================================
Total params: 5,298,226
Trainable params: 5,298,226
Non-trainable params: 0
```

Figure 9: Detailed Result

VI. CONCLUSION

This paper has proposed acoustic scene classification with DenseNet Algorithm which provides us a decent achievable accuracy on ESC – 50 dataset. The paper deals with preprocessing and transformation of datasets through data augmentation via adding Gaussian white noise for practical observation.

This paper initialized with standard audio format files which we analyzed through the characteristics of audio like Mel-spectrogram, chroma, Tonnetz, and MFCC. Further, the Mel-Spectrogram was transformed, augmented, and fed as an input in our DesneNet Model. Finally, the classification of the environmental sounds has achieved 96% accuracy on dataset with 86.43% as validation accuracy with 5% of white noise.

VII. REFERENCES

[1] Spoorthy. V, Manjunath Mulimani – "Acoustic Scene Classification using Deep Learning Architectures"

[2] Bregman, A. S. – " Auditory scene analysis: The perceptual organization of sound."

[3] Brown, Guy Jason – "Computational auditory scene analysis : a representational approach"

[4] Wen Zhao; Bo Yin – "Environmental sound classification based on adding noise"

[5] Wei Qin, Bo Yin – "Environmental Sound Classification Algorithm Based on Adaptive Data Padding"

[6] R. Radhakrishnan, A. Divakaran and A. Smaragdis, -"Audio analysis for surveillance applications,"

[7] Behnaz Bahmei- "CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification" IEEE signal processing (vol 29)

[8] Wenjie-Mu – "Environmental sound classification using temporal – frequency attention based convolutional neural network"

[9] Manjunath Mulimani – "Acoustic Scene classification using Deep Learning Architectures"

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition,"

[11] Daniele Barchiesi and Dan Stowell – "Acoustic Scene Classification"

[12] Jianrui Lu, Ruofei Ma, Gongliang Liu
- "Deep Convolutional Neural Network with Transfer Learning for Environmental Sound Classification"

[13] Lidong Yang, Jiangtao Hu – "Research on Acoustic Scene Classification Based on Multiple Mixed Convolutional Neural Networks": 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)