e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-6 (December 2021)
https://doi.org/10.31033/ijemr.11.6.17

Environmental Audio Tagging Using Deep Convolution Neural Network and Digital Signal Processing

Anirudh Rana[1] and Rajinder Singh Rana[2]
[1]System Security Engineer, Silex Software's Ltd, INDIA
[2]Principal, Sanatan Dharama College, Ambala Cantt., INDIA

[1]Corresponding Author: anirudhrana3976@gmail.com

ABSTRACT

Machine learning has experienced a strong growth in recent years, due to increased dataset sizes and computational power, and to advances in deep learning methods that can learn to make predictions in extremely non-linear problem settings. The intense problem of automatic environmental sound classification has received alarming attention from the research community in recent years. In this paper the audio dataset is converted into mass spectrogram using Digital Signal Processing (DSP). The spectrogram thus obtained is fed to the Convolutional Neural Network (CNN) for the classification of the audio signal. In this we present a deep convolutional neural network architecture with localized kernels for environmental sound. By training the network on another additional deformed data, the hope is that the network becomes invariant to all deformations and generalizes better to all unseen data. We show that the proposed DSP in combination with CNN architecture, yields state-of-the-art performance for environmental sound classification.

***Keywords*--** Environmental Sound Classification, Deep Convolutional Neural Networks, Digital Signal Processing, Urban Sound Dataset, Data Augmentation

I. INTRODUCTION

Machine Learning has experienced a very strong growth in recent years, due to increased dataset sizes and computational power, and to advances in deep learning method that can learn to make predictions in extremely nonlinear problem setting.

A lot of research has been done and is in process in tagging of audio recordings. In [1] the authors proposed a content based automatic sound tagging scheme using the deep convolutional neural network. As seem with most of the tasks, the first step is always to extract features from the audio sample. Then sorting is done according to the nuances of the audio. This can be done by machine learning or deep learning. In [2], the author used conventional Network to perform environment sound tagging and tagging the presence of any desired element in the audio.

Depending upon the audio event to be detected and classified in each task, it may become difficult to collect enough samples for them. Furthermore, different task use task specific dataset hence the amount of recording may be limited. In [3], the report emphasized on the weakly labelled data that takes much less time, since the annotator only mark the active sound event classes and not there boundaries.

In [4] the author used a joint detection classification network that slices the audio into blocks, audio detector and classification on each block then uses the overall audio tag to train using a weak label of recording.

II. METHODLOGY

For performing the data science on audio signal, we first need to convert the audio signal into a proper format that can be implemented through the Neural Network. This process of converting the audio signal into the desired format is called Featurization and the process applied in the following case for featurization of audio sample is Digital Signal Processing.

A. Digital Signal Processing

The input audio sample that we recieve are in the range of minutes. Thus, owing to this condition we cannot perform the processing task of audio samples therefore we divide the given Audio samples into the many many small subframes. The sample is divided into subframes of 30-40 msec. These subframes are so minute in size that they appear to be stationary in nature. And we require the same stationary element for the making of the periodogram. This process of Audio sample division and designing of the periodogram through these stationary elements are called " Fast Fourier Transform ".

After the completion of Fast Fourier Transform process, the Mass Spectrogram is obtained from each periodogram formed from each uniquely divided subframes.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-6 (December 2021)
https://doi.org/10.31033/ijemr.11.6.17

A Spectrogram is basically a heat signature that is represented pictographically. Spectrogram helps us evaluate the variation of frequency with respect to time.

This entire process featuring the Fast Fourier Transform and mass spectrogram is called "Audio Fingerprinting ". Each audio has its own unique fingerprint. These fingerprints are the ones responsible in predicting which tag is to be applied on the given audio sample. The audio fingerprints obtained is fed to the CNN network for training purpose.

Data Augmentation is a way of creating new data with different orientations of the same data. The benefits of this are two folds, the first being the ability to generate ‘more data’ from limited data and secondly it prevents over fitting. Each deformation in data is applied directly to the audio signal prior to converting it into the input representation used to train the network (log-mel-spectrogram). The deformations and resulting augmentation sets are given below:

- Time Stretching (TS)
- Pitch Shifting (PS1)
- Pitch Shifting (PS2)
- Dynamic Range Compression (DRC)
- Background Nube (BG)

The most common acoustic features used to represent spectral content of audio signals are mel-band energies and mel-frequency cepstral coefficients (MFCCs). Their design is based on the observation that human auditory perception focuses only on magnitudes of frequency components. The perception of these magnitudes is highly non-linear, and, in addition, perception of frequencies is also non-linear. Following perception, these acoustic feature extraction techniques use non-linear representation for magnitudes (power spectra and logarithm) and non-linear frequency scaling (mel-frequency scaling). The non-linear frequency scaling is implemented using filter banks which integrate the spectrum at non-linearly spaced frequency ranges, withnarrow band-pass filters at low frequencies and with larger bandwidth at higher frequencies.

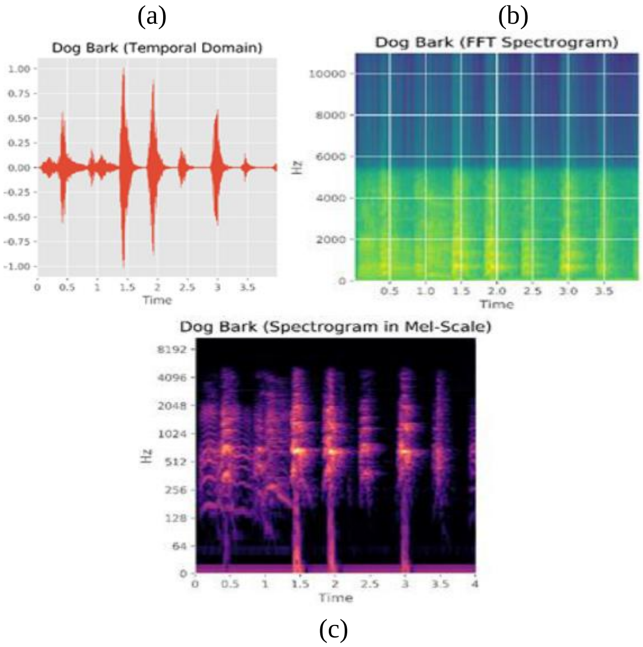


Figure 1: (a) Dog Bark Audio Signal (b) Dog Bark Spectrogram obtained after applying FFT (c) Dog Bark Spectrogram obtained in Mel-Scale

B. Convolutional Neural Network

A Convolutional Neural Network (CNN) is a class of deep neural network, most commonly is applied for analysing visual imagery. CNN works upon the concept of multilayer perceptrons designed to require minimal preprocessing.

Convolution networks were inspired by biological process in that the connectivity pattern between neurons resembles the organization of human system.

The audio finger print that were obtained from digital signal processing is passed through the CNN.

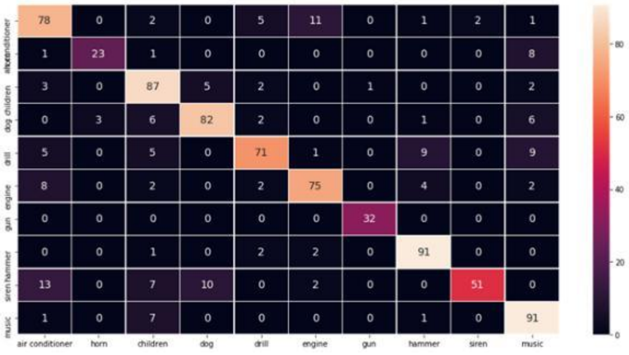
The audio fingerprint obtained are passed through three neural networks namely:-

- Convolutional
- Pulling
- Flattening

The convolution layer divides the spectrogram into smaller subparts. Pulling layer (max pulling), since it extracts the maximum frequency present in the small subparts and by this virtue, a multi-dimensional array is obtained. Flattening converts the multi-dimensional array obtained from pulling into a 1-D array. As each frequency has its own unique 1-D array so the 1-D array obtained from the sample is compared with the stored frequency 1-D array, accordingly the tag is given to the audio.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-6 (December 2021)
https://doi.org/10.31033/ijemr.11.6.17

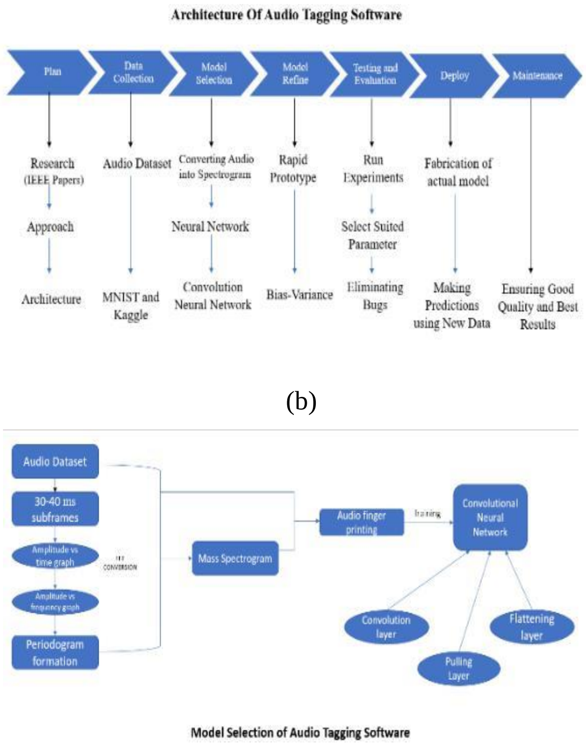
After training the Machine through CNN. The software becomes trained and the machine becomes ready to test and work.



	air conditioner	horn	children	dog	drill	engine	gun	hammer	siren	music
air conditioner	78	0	2	0	5	11	0	1	2	1
horn	1	23	1	0	0	0	0	0	0	8
children	3	0	87	5	2	0	1	0	0	2
dog	0	3	6	82	2	0	0	1	0	6
drill	5	0	5	0	71	1	0	9	0	9
engine	8	0	2	0	2	75	0	4	0	2
gun	0	0	0	0	0	0	32	0	0	0
hammer	0	0	1	0	2	2	0	91	0	0
siren	13	0	7	10	0	2	0	0	51	0
music	1	0	7	0	0	0	0	1	0	91

Figure 2: Test set confusion matrix along with the resultant 1-D array obtained

(a)



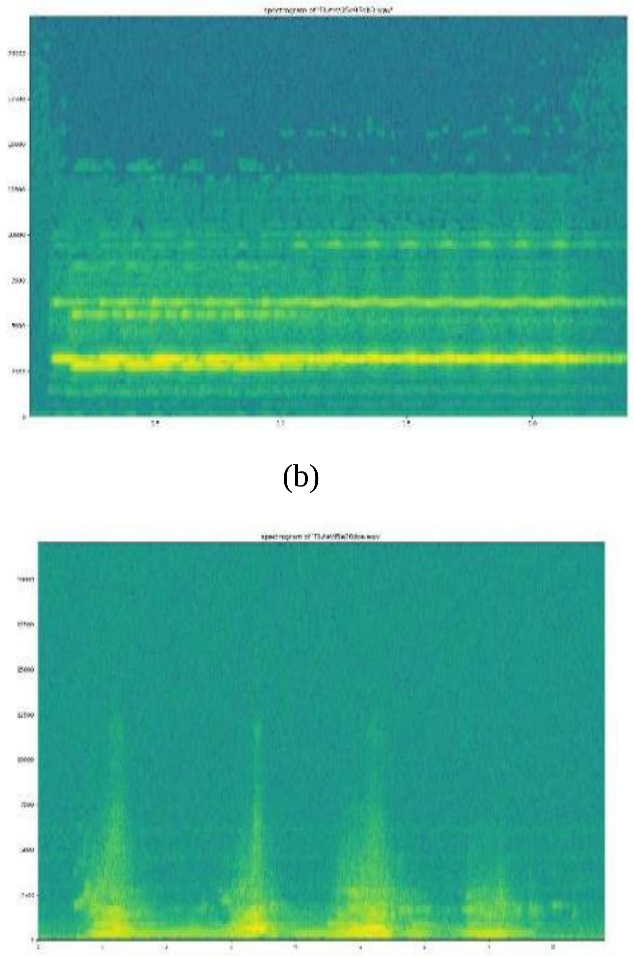
(b)

Figure 3: (a) Architectural design of our model (b) Model Selection of our software

IV. RESULTS

The codes were executed and favourable results were obtained. We were successfully able to convert all the audio sample into required mass spectrograms.

(a)



(b)

Figure 3: (a) and (b) are the obtained spectrogram after applying FFT

V. CONCLUSION

In this article we proposed a digital signal processing and deep convolutional neural network with a set of audio data. We showed that the improved performance stems from the combination of DSP and deep CNN together.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-6 (December 2021)
https://doi.org/10.31033/ijemr.11.6.17

REFERENCES

[1] ML Blog Team. (2021). *Hearing AI: Getting started with Deep Learning for Audio on Azure*.

[2] Justin Salamon & Jaun Pablo Bell. (2015). *Deep convolution network and data augmentation for environment sound classification*.

[3] Guojun Lu & Templar Hankinson. (1998). *A technique towards automatic audio classification and retrieval*. Available at: Monash University, Australia.

[4] Tomi Heittola & Tuomas Virtanen. (2018). *The machine learning approach for analysis of sound scenes and events*.