ISSN:2583-3286(Online)

IPEC Journal of Science & Technology
Vol. 01 (01), Feb. 2022, pp. 7-10

Audio Classification Using Deep Learning

Kirti Jain*, Umang Taneja*, Deepesh Singh Tomar*, Mohit kashyap*, Aman*

*Department of Computer Science and Engineering,
Inderprastha Engineering College, Ghaziabad, India*

Accepted on 23 February 2022

Abstract: Sound is an inevitable signifier that provides us relevant information related to the environment around us. It helps us perceive the environment and our ears can distinguish real world audio signals that are overlapped with one another easily. However, it is difficult to replicate human ear like functionality to recognize each sound signal distinctly with the standard detection methods. With the help of a state-of-art CNN architecture we will classify the sound using spectrogram-based inputs such as mel spectrogram. After that, we developed a mobile application to ease the process of taking the audio signal from the user and output the result in the application itself.

Keywords **- Audio Classification, Deep learning, Sound, Feature Extraction, Signal Processing.**

.

I. INTRODUCTION

Sound is an environmental signifier that helps us to collect relevant information related to the environment we live in. Our brain continuously processes and understands the audio data and helps us to extract information related to the environment. Sound is essentially an audio signal that is received by our ears and perceived by our brain. It consists of two attributes- amplitude and frequency. It is a wave-like format where the amplitude and frequency changes with respect to time.

Audio signal processing has become one of the highly researched fields due to its wide range of applications. Audio sensing enables a user to detect his daily activities like driving to the office, working on the computer, having lunch etc. There are a variety of sound processing and machine learning techniques that have been applied but the most effective one to use is a state-of-art CNN. The CNN is predominantly used in Image classification but advancements in Image classifications have made it possible to use CNN for sound classification purposes. By converting the audio signal to frequency domain and extracting the features from it, we can easily generate the spectrograms. These spectrograms are provided as input to the CNN for sound classification.

Sound Event classification is a prominent area of research with it having applications in a variety of areas, ranging from security surveillance, wildlife monitoring and healthcare. Researchers in the advanced deep learning field were majorly focused on single sound event classification. The problem of Audio classification can be solved using deep learning models.

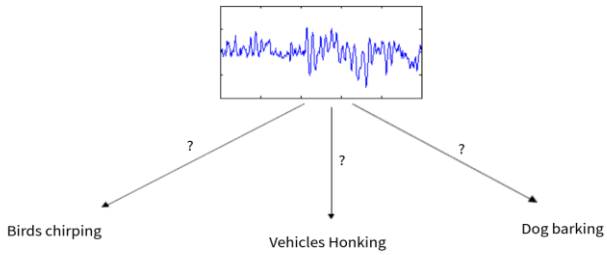

Figure 1: Audio classification in a nutshell

In short, the audio classification is the task of taking the audio sample as input and determining the class from it. Example: from a given audio signal find if it has a bird chirping, vehicle honking and dog barking.

Polyphonic detection in the terminology of audio signal preprocessing detects and classifies each of the sound events from the mixed audio signal. But, it is an unconstrained real-world task that is yet to be explored.

In our research instead of directly using the amplitude vs time audio signal as input, we will do some processing by applying the Short Time Fourier Transform to the audio signal. The STFT, involves the splitting of an audio signal into frames and then taking Fourier transform of each frame. The Fourier is a modern way to decompose an audio signal into its constituent frequencies in audio processing.

Date of Submission: 18 Jan 2022
Coresponding Author: Kirti Jain (e-mail: kirti.jain@ipec.org.in).

Published online : 25 February 2022

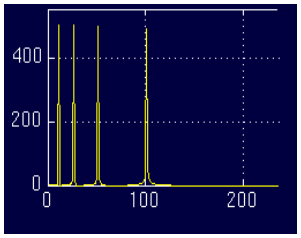

Figure 2: Time domain input signal

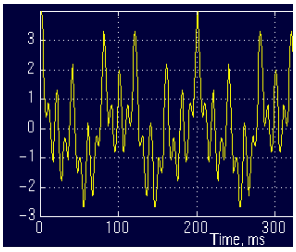

Figure 3: Output of conversion

We have used Mel Frequency Cepstral Coefficients. Figure 3 shows how the MFCC features will be extracted from the audio signal.

Followed by a Spectrogram based end-to-end image classification using a CNN pre-trained on AudioSet to learn acoustic features, a binary classifier to characterize each sound category, and a classifier to learn the outcome of each binary classifier for multi-sound classification.

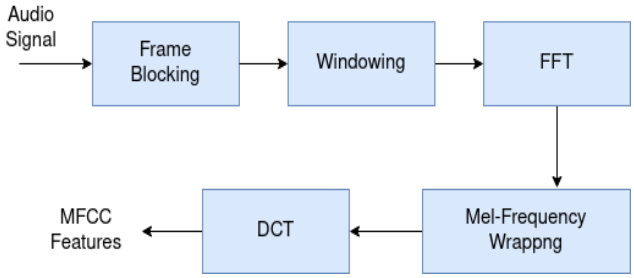

Figure 4: Process of calculating MFCCs features from the audio signal.

There are many use cases of Sound event classification but few of them are mentioned below:

- Baby monitoring
- Fire alarm detection for hearing impaired people
- Engine sound analysis for maintenance etc
- Chatbots and virtual assistant
- Machine translation and text to speech applications.

Google launched a dataset called the AudioSet - which is a large collection of labeled audio taken from YouTube videos which we will be using for our use case. The workflow of the research is given in figure 4.

II. RELATED WORK

Some interesting properties of deep neural networks are periodically showing up in the scientific literature. Their ability to solve audio classification problems have led to the rise of interest of researchers in the audio classification field. Their work explored many methods in the field of ML that were suitable for single audio classification. For example- Support vector machines with continuous wavelet-transform were able to produce great results.

A state-of-art CNN based convolutional neural network can be used to enhance the accuracy and efficiency of the model. A CNN based model can be trained for several individual audio classification tasks like- music genre classification, gender classification etc.

In [2], the authors trained two datasets- ESC-10 and ESC-50 and achieved accuracy as 77% and 49% in CNN. It can be concluded that by using the spectrogram images of sounds, the problem of sound classification and recognition systems can be solved.

The research also deep dive into audio classifiers using Artificial neural networks by segmenting and characterizing them into different categories of audio. But, due to a fairly small amount of data the accuracy wasn't the best. Also, use of CNN has proved to be more efficient than using an artificial neural network.

Most of the work that has been done prior is related to the monophonic detection i.e. detect a single sound source and classify it based on the classes. The polyphonic detection is challenging and yet to be explored. The most common approach to extract the frequency domain features is to use MFCCs. The MFCCs help the model to replicate human-like behaviour. It is one of the most popular steps amongst audio classification tasks to extract features by using MFCCs and generating log mel spectrograms.

III. METHODOLOGY

We proposed an approach that would classify multiple sounds in an audio signal without prior knowing how many signals are mixed together. The purpose of this approach is to classify multiple sound events from real world audio as the real world audio signal is overlapped with one or more than one audio class. The approach is built around the state-of-art CNN model, which is predominantly used for the image classification related tasks but we will use it for our audio classification model as well. We developed a stacked classifier to detect each and every audio class from the given audio signal without prior knowing the number of classes that are

overlapped with each other. Each phase is explained briefly below:

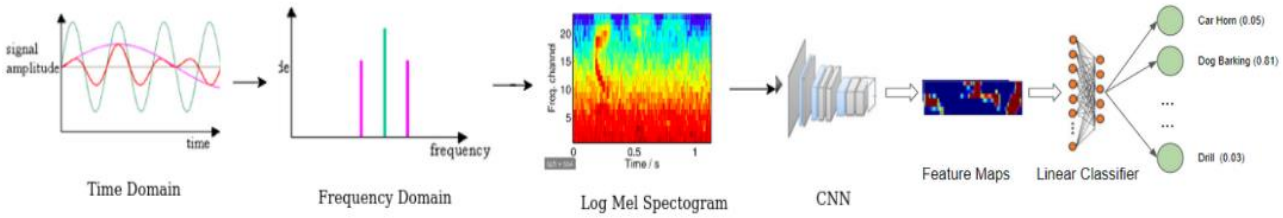

Figure 5: Workflow of the overall research

Pre-processing

Data preprocessing involves the following steps:

- Read the input signal
- Resample it
- Convert to mono channel
- Remove the unwanted noise
- Trim the audio

The purpose of preprocessing the audio signal is to standardize all the audio signals and to prepare them for feature extraction. The first step is to read the real world audio from the user. After having the input audio signal, the next step is to perform data preprocessing. We resampled the given audio signal to 20Khz. 44100Hz is the most popular sampling rate but we resampled it to 20KHz as the highest frequency that can be heard by the human ears is 20Khz. The next step in the data preprocessing is to mono audio signal from stereo signal for simplicity. Since the real world data consists of noise there is a need to preprocess the data before we can extract features from it. We would remove the unwanted noise from the audio signal using one of the libraries. After removing the unwanted noise, we will then trim the parts of the audio that are not needed for feature extraction.

Feature Extraction

After data preprocessing, we then extracted features from each frame of the given audio signal. The feature extraction is an important step for classifying an audio as it helps in analysing the audio signal and finding possible relations among signals. There are many approaches for feature extraction. We used MFCC to extract features from each frame of the given audio signal by computing the spectrogram by first converting the time domain input signal to frequency domain using STFT. We were able to extract the log Mel spectrogram by taking the logarithm of the spectrogram that was computed earlier.

Steps that were performed in this phase are:

- A audio clip is further divided into frames, which are the smallest units on which feature identification is possible
- Applying a Fourier Transformation provides us with a summary of the frequency distribution over a frame
- Mel-frequency cepstral coefficients (MFCCs) are values which characterize a spectrum by its spectral envelope, but non-linearly (as opposed to a standard linear cepstrum, which is provided by an FT)
- In order to derive the MFCCs, a logarithmic mapping function is applied which divides the distribution into bands associated with each MFCC
- This is in order to more closely approximate the perception of the human ear, to which frequency response is logarithmic rather than linear
- It also helps to make the classifiable features of the clip more prominent, which later can be used for audio classification tasks.

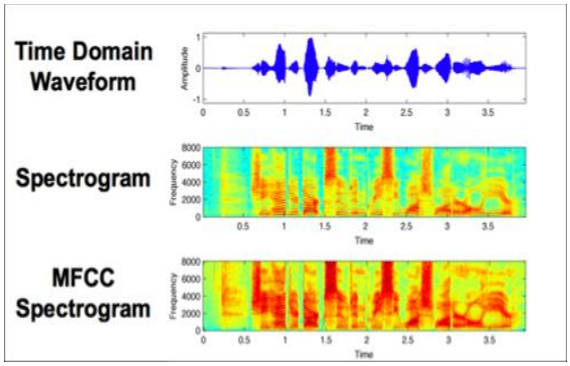

Figure 6: Spectrograms and MFCC Spectrogram

Pretrained CNN

After computing features, we provided these extracted features to a pretrained CNN. We used YAMNet CNN for the sound classification. The YAMNet CNN is often used for sound classification and it is pretrained on Audioset dataset. It can predict over 521 classes from the Audio Set-Youtube corpus. The last layer of the CNN

Published online : 25 February 2022

consists of an embedding layer 128 wide fully connected layer.

Mobile application

We finally created a mobile application to be able to solve the audio classification problem using deep learning techniques using real world dataset. So, the user can directly use his mobile phone's microphone to input the audio signal and the model will classify into one of the different categories. And finally output the result on the screen for the user to see.

IV. RESULTS AND DISCUSSION

We developed a mobile application for the audio classification of real world audio signals that will be input by the user from their phone's microphone. The audio signal received from the user is then passed through a series of phases before the final classification part. We were able to resample, remove noise and trim the input audio signal for standardization and preprocessed it before extracting features from it. The extracted features are then passed onto a pretrained Convolutional neural network which eases the process of classification by classifying the audio signal into one of the known classes. The output is then displayed on the screen to show the user with the category along with the percentage of that category found in the given audio signal.

V. CONCLUSION

The mobile application contains the functionality of using the microphone of the mobile to get the input audio from the user. After having the input audio, clean it and preprocess it before going for the feature extraction phase. In feature extraction phase, generate the log mel-spectrogram of the input audio signal and use these spectrograms as input for the pretrained Convolutional Neural Network YAMNET, and classify the input audio signal into one of the classes and show it as output to the user within the mobile application. Thus, the application can be used to classify the audio signal using deep learning techniques. This classification can be further used to get all the possible classes in the audio signal as it is very likely that real world audio contains overlapping acoustic signals and it is not easy to detect all possible classes with traditional techniques.

REFERENCES

[1] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen. Polyphonic sound event detection using multi-label deep neural networks. In IJCNN '15, pages 1–7, July 2015.

[2] Karthikeyan & Mala (2018). CONTENT BASED AUDIO CLASSIFIER & FEATURE EXTRACTION USING ANN TECHNIQUES.IJIRAE::International Journal of Innovative Research in Advanced Engineering, Volume V, 106-116.

[3] J. F. G. et al. Audio set: An ontology and human-labeled dataset for audio events. In ICASSP '17, New Orleans, LA, 2017.

[4] C. Ghita, R. D. Raicu, and B. Pantelimon. Implementation of the fast ICA algorithm in sound source separation. In ATEE '15, pages 19–22, May 2015.

[5] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda. Duration-controlled lstm for polyphonic sound event detection. IEEE/ACM Trans. Audio, Speech, and Lang. Proc., 25(11):2059–2070, Nov. 2017.

[6] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using a source-filter model for sound separation. In ISMIR, 2009.

[7] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C.Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In ICASSP '17, pages 131–135. IEEE, 2017.

[8] P. Jincahitra. Polyphonic instrument identification using independent subspace analysis. In ICME '04, 2004.

[9] N. D. Lane, P. Georgiev, and L. Qendro. Deepear: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In UbiComp '15, pages 283–294, 2015.

[10] A. Mesaros, T. Heittola, and T. Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6, 2016.

[11] G. Parascandolo, H. Huttunen, and T. Virtanen.Recurrent neural networks for polyphonic sound event detection in real-life recordings. In ICASSP '16, 2016.

[12] Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3):279–283, 2016.