Themed Section : Engineering and Technology
© 2018 IJSRSET | Volume 5 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099

Deep Learning based Bird Audio Detection

E. Sophiya[1], S. Jothilakshmi[2]

[1]Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamilnadu, India

[2]Department of Information Technology, Annamalai University, Annamalainagar, Tamilnadu, India

venus.sophiya@gmail.com [1], jothi.sekar@gmail.com [2]

ABSTRACT

Audio event detection (AED) is defined as analyzing a continuous acoustic signal in order to extract the sound events present in the acoustic scene. Sound events are best labels for an auditory scene, because they help in describing and understanding a recognizable event present in the sound. In this work, Bird audio detection is carried out to determine whether birds sound is present in the given environmental audio. The system is designed with machine learning algorithm using Tensorflow. The proposed model learns spectrogram features from audio and predicts the presence of bird sounds with an accuracy of 80.76%.

Keywords: Audio processing, Audio scene analysis, Audio event detection, Bird audio detection, Deep learning, Tensorflow, Audio features.

I. INTRODUCTION

Automatic detection of animal vocalizations, such as singing birds, has a scientific challenge in itself, can be helpful for monitoring biodiversity. Various researches are involved in bioacoustics, the recent field named ecoacoustics that gathers ever-growing quantities of recordings that need to be manually analyzed. Automatic tools that label the recordings accurately are very much in demand to ease the time-consuming task of listening to hours of them [1]. Monitoring birds by their sound is important for many environmental and scientific purposes. A variety of crowdsourcing and remote-monitoring projects now record these sounds, and analyze the sound automatically. Still there are many issues to solve, as indicated by the number of projects that are yet to be fully automated.

Bird audio detection is one of intensive research given by Detection and Classification of Acoustic Scenes and Events (DCASE 2018), due to the fact that birds are more easily detectable through the audio modality rather than vision. Bird audio detection (BAD) aims to detect the presence or absence of bird calls in an audio recording. BAD is an Audio event detection (AED) task, which aims to identify all the audio events and their occurrence time in a mixed audio recording. BAD have many applications in environmental science, such as monitoring the density and the migration of birds in depopulated zone. Bird audio detection (BAD) is defined as identifying the presence of bird sounds in a given audio recording. Figure 1 shows the Bird audio detection system.

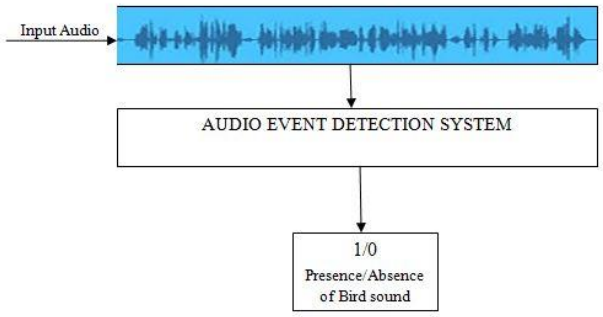

Figure 1. Audio Event Detection System for Birds Sound

In many conventional, remote wildlife-monitoring projects, the event detection process is not fully automated and requires heavy manual labor to label the obtained data [2], [3]. In certain cases such as dense forests and low illumination, automated detection of birds in wildlife can be more effective through their sounds compared to visual cues. The challenge provides three bird audio datasets recorded in different acoustic environments. Bird sounds can be broadly categorized as vocal and non-vocal sounds. In this research work, since non-vocal bird sounds are harder to associate with birds without any visual cues, the research on BAD has been mostly focused on vocal sounds. The challenge provided annotated and non-annotated bird call recordings. The former is utilized as the training dataset and the latter are recordings from a completely different geographical location and employed as the test dataset.

Different locations also mean different acoustic environments leading to a variety of sound sources. Furthermore, each of these bird species has unique calls, resulting in a wide variety of bird calls. Labeling such a wide variety of calls into one class weakens the classifier and can result in misclassification of similar sounding non-bird sounds. The problem is further intensified in the dataset used because each of the bird calls has been recorded with different devices that add their own system noise. A bird audio detection method which can work across such a wide range of species and environments is termed as a generic method. BAD have many difficulties.

First, in an audio clip only the presence or the absence of birds is known but not knowing the time stamps of the bird calls and other sounds. Detecting the presence of bird calls in audio recordings can serve as a basic step for wildlife and biodiversity monitoring. To help advance the state of the art in automating this task, Bird audio detection challenges were organized by DCASE. Specifically, participants were asked to build algorithms that predict whether a given 10-second recording contains any type of bird vocalization, regardless of the species. Use of passive acoustic monitoring to estimate animal population density and there is increasing interest in using acoustic indices for biodiversity assessments.

In this work, the bird audio detection using simple convolutional neural networks (CNNs) is proposed. The CNN architecture exploits the combined modeling of fully connected four layers. CNN architecture has also been proposed in automatic speech recognition and music classification [4]. Bird species identification has been the target of several international evaluation campaigns such as LifeCLEF (BirdCLEF), a yearly contest including bird species identification in various audio recordings. A variety of machine learning techniques have been explored for this task. The winning solutions of last year BirdCLEF edition were based on deep convolutional neural networks (CNNs) [5]. Indeed, the success of deep learning (DL) and deep neural networks (DNN) in many domains involving classification tasks, offers new and appealing perspectives. In the context of the BAD challenge, concerned with the detection of bird sound in short duration recordings, the evaluation reported is based on CNNs, more specifically on densely connected CNNs, also called denseNets [6].

The rest of the paper is organized as follows. The related work of bird audio detection is reviewed in Section 2. The proposed framework is explained in

Section 3. Experimental results obtained by proposed method are discussed in Section 4. Finally, Section 5 concludes the paper.

II. Related Works

Sharath Adavanne, Konstantinos Drossos, et al [7] proposed the novel method for detection of bird calls in audio segments using stacked convolutional and recurrent neural networks. Data augmentation using blocks mixing and domain adaptation using the test mixing is implemented to make the method robust for unseen data. Thomas Pellegrini [1] proposed to detect bird sounds in audio recordings automatically and monitoring biodiversity based on audio field recordings. The author experimented several types of convolutional neural networks to estimate how accurate the state-of-the-art machine learning approaches and the bird audio detection challenges are reported in the framework. Emre Cakir, Sharath Adavanne, et al [8] proposed the convolutional recurrent neural networks on the task of automated bird audio detection in real-life environments. The convolutional layers extract high dimensional, local frequency shift invariant features, while recurrent layers capture longer term dependencies between the features extracted from short time frames. Qiuqiang Kong, Yong Xu, et al [9] implemented Bird audio detection (BAD) to detect whether there is a bird call in an audio recording or not. They applied joint detection and classification (JDC) model on the weakly labeled data (WLD) to detect and classify an audio clip at the same time. Anshul Thakur, R. Jyothi, et al [10] describes the bird activity detector which utilises a support vector machine (SVM) with a dynamic kernel. Bird activity detection is the task of determining if a bird sound is present in a given audio recording. Dynamic kernels are used to process sets of feature vectors having different cardinalities. Probabilistic sequence kernel (PSK) is one such dynamic kernel. Thomas Grill and Jan Schlüter [11] present and compare two approaches to detect the presence of bird calls in audio recordings using convolutiona neural networks on mel spectrograms. Comparing multiple variations of the systems, find that despite very different architectures, both approaches can be tuned to perform equally well. Dan Stowell, Mike Wood, et al [12] reviews the state of the art in automatic bird sound detection, and identifies a widespread need for tuning-free and species-agnostic approaches.

III. Proposed Work

Deep learning refers to a machine learning techniques that uses supervised or unsupervised algorithm to automatically learn multiple levels of representations in deep architectures. Inspired by the human biological neurons for processing natural signals, deep learning has reached much attention in various research domains such as speech recognition, computer vision, and image analysis. Figure 2 shows Deep Neural Network (DNN) architecture.

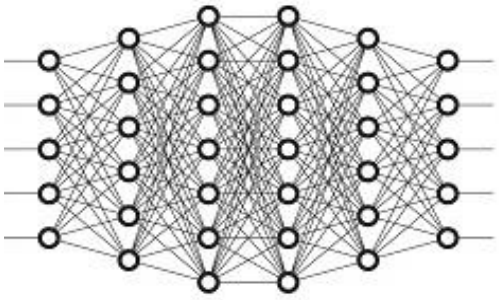

Figure 2. Architecture of DNN

Tensorflow is used to implement complex DNN structures without getting complex mathematical details, and availability of large datasets. Tensorflow has a high-level machine learning API which makes easy to configure, train, and evaluate a large number of machine learning models. Keras, a high-level deep learning library, is used on top of Tensorflow. It supports various DNNs like RNNs, CNNs, and a combination of the two architectures.

In the proposed work spectrogram features are learned from the audio signals to detect the bird audio.

The machine learning model is designed with CNN with three hidden layers. This network is implemented in keras. In this system, the official dataset of IEEE AASP Challenge (2018) on Detection and Classification of Acoustic Scenes and Events (DCASE) is used. The dataset consists of 10sec long wav files with 44.1 khz mono PCM. The data files are manually labeled with a 1 and 0 to indicate the presence/absence of any birds audio within that audio clip.

For each input audio signal, Short Term Fourier Transform (STFT) magnitude spectrogram with a window size of 10ms is extracted over 40 ms audio frames of 50% overlap, windowed with hamming window. Librosa library is used in the feature extraction process.

The spectrogram is a representation of the audio signal in time-frequency domain. The relevant information content of spectrogram will be high in a time domain of a signal. Thus the proposed work can easily detect the bird audio event within an audio clip. Figure 3 and Figure 4 shows the visual representation and spectrogram representation of audio with and without birds sound. The preprocessed data is then mapped to the training and test data. The dataset has audio files with and without bird sounds. So applying normalization will improve the prediction of events within the audio. The features are then normalized using scikit learn preprocessing function.

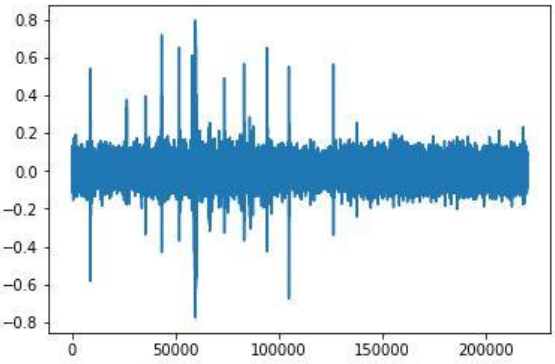

Figure 3 (a). Representation of Audio with Bird sound

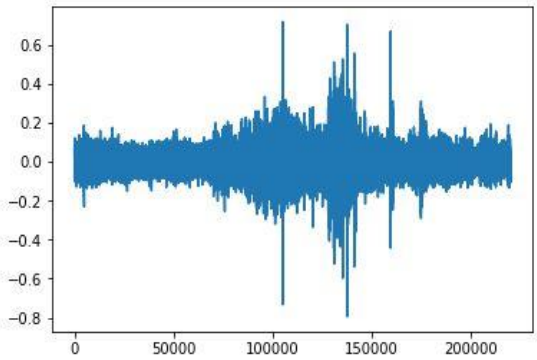

Figure 3 (b). Representation of Audio without Bird sound

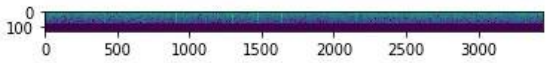

Figure 4 (a). Spectrogram of Audio with Bird sound

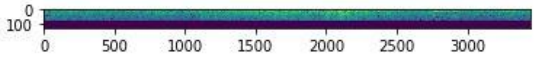

Figure 4 (b). Spectrogram of Audio without Bird sound

The preprocessed data is then mapped to the training and test data. The dataset has audio files with and without bird sounds. So applying normalization will improve the prediction of events within the audio. The features are then normalized using scikit learn preprocessing function. The machine learning model used in this work is CNN and the network is implemented with keras. The keras library will support best function to try various network architectures. The architecture of model consists of four dense layers with 128, 32, 32, and 2 units that detect the bird audio. Each convolution layer is mapped with Relu activation function and output layer is mapped by softmax function.

IV. Experimental work and Results

The objective of the feature learning in the network is to get the estimated bird audio event closer to binary target outputs, where target output is 1 if any bird sound is present in a given recording and 0 vice versa.

For the prediction of birds audio, the dataset is randomly partitioned into 60% for Training, 20% for

Validation, 20% for Testing. The proposed model is evaluated using three cross fold validation scheme.

TABLE I
Development dataset for Bird audio detection

Dataset	Bird Sound	
	Present	Absent
freefield1010	5755	1935

The network is trained with back-propagation through time using Adam optimizer and binary cross-entropy as the loss function. Keras deep learning library has been used to implement the network. The scores for validation and test dataset are presented in Table 2. The proposed network outputs a probability of 80.76% for birds sound in these recordings.

TABLE III
SCORES FOR VALIDATION AND TEST DATA WITH SPECTROGRAM FEATURE

Feature	Accuracy (%)	
	Validation	Test
Spectrogram	90.9	80.76

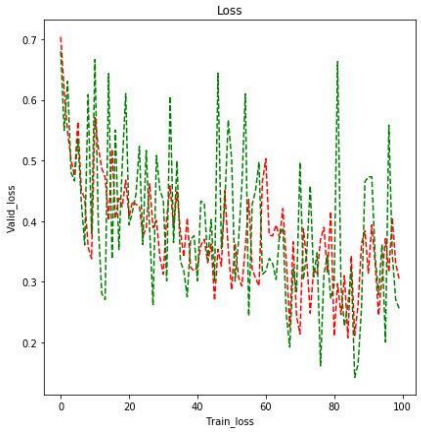

Figure 5(a). Loss

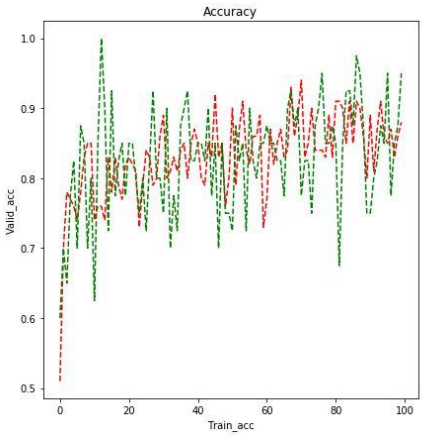

Figure 5(b). Accuracy

V. CONCLUSION

In this work a bird audio detection is carried out based on deep learning algorithm using Tensorflow. The objective of the model is to estimate the presence/ absence of birds sound in a given audio. The audio detection system outputs a zero if none of the target species are detected, and as one otherwise. The datasets are collected from DCASE 2018 challenge. CNN network with spectrogram features were used to model the system. The proposed work thus predicts the bird audio by achieving an accuracy of over 80.76%. In future, this work will be extended with different network architecture and the model can also be hyper tuned with parameters to predict the events better with higher accuracy.

VI. REFERENCES

[1] Thomas Pellegrini "Densely Connected CNNs for Bird Audio Detection" 25th European Signal Processing Conference (EUSIPCO), 2017.

[2] R. T. Buxton and I. L. Jones, "Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration," Journal of Field Ornithology, vol. 83, no. 1, pp. 47–60, 2012.

[3] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, "Estimating animal population density using

passive acoustics," Biological Reviews, vol. 88, no. 2, pp. 287–309, 2013.

[4] E. C. akır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in IEEE/ACM TASLP Special Issue on Sound Scene and Event Analysis, 2017

[5] E. Sprengel, Y. Martin Jaggi, and T. Hofmann, "Audio based bird species identification using deep learning techniques," Working notes of CLEF, 2016.

[6] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," arXiv preprint arXiv:1608.06993, 2016.

[7] Sharath Adavanne, Konstantinos Drossos, Emre Cakır, Tuomas Virtanen "Stacked Convolutional and Recurrent Neural Networks for Bird Audio Detection" 25th European Signal Processing Conference (EUSIPCO), 2017

[8] Emre Cakir, Sharath Adavanne, Giambattista Parascandolo, Konstantinos Drossos and Tuomas Virtanen "Convolutional Recurrent Neural Networks for Bird Audio Detection" 25th European Signal Processing Conference (EUSIPCO), 2017

[9] Qiuqiang Kong, Yong Xu, Mark D. Plumbley "Joint Detection and Classification Convolutional Neural Network on Weakly Labelled Bird Audio Detection" 25th European Signal Processing Conference (EUSIPCO), 2017

[10] Anshul Thakur, R. Jyothi, Padmanabhan Rajan, A.D. Dileep "Rapid Bird Activity Detection Using Probabilistic Sequence Kernels" 25th European Signal Processing Conference (EUSIPCO), 2017

[11] Thomas Grill and Jan Schlüter "Two Convolutional Neural Networks for Bird Detection in Audio Signals" 25th European Signal Processing Conference (EUSIPCO), 2017

[12] Dan Stowell, Mike Wood, Yannis Stylianou and Herve Glotin "Bird Detection In Audio: A Survey and a Challenge" IEEE International Workshop On Machine Learning For Signal Processing, Salerno, Italy, Sept. 13–16, 2016