Detecting Bird Vocalizations in Audio Files With a Convolutional Neural Network

1st Wisdom Aduah
African Institute for Mathematical Sciences
Muizenberg, South Africa
wizzy@aims.ac.za

***Abstract*—In this study, we train a Convolutional Neural network classifier to detect bird vocalizations in an audio file. Bird sounds were recorded at Intaka Island Nature Reserve using Raspbery Pi. The model achieved an accuracy of 78.3% and a precision of 81.1%.**

***Index Terms*—Deep learning, vocalizations, Bioacoustics, ARU, CNNs**

I. Introduction

Deep Learning has had tremendous success in learning images, language and even audio. However, it requires massive amounts of data to train [1]. Animals produce distinct vocalizations, which enables them to be identified simply by their sound. In this study however, we will only work on distinguishing bird vocalizations from non-bird vocalizations.

Automatic Recording Units (ARUs) were used in this project. Unlike cameras, audio recorders can capture signals (in this case sounds) that are not directly in front of them; they can capture vocalizations of animals that are well hidden in their habitats which do not easily interact with the environments. ARUs are programmable units that automatically record sound in the field on a set schedule. They are ideal for remote observations, and allow researchers spend less time in the field when they make visits. This makes them very useful for targeting rare species. They also provide a permanent record of the data. [2]. Despite their usefulness, they have the following drawbacks: can be expensive depending on the setup. Also, hardware failures could lead to loss of data; storage could be an issue, especially when recording at high sampling rates.

II. Methodology

A. Data collection and annotation

The data for the study was recorded at Intaka Island Nature Reserve in Cape Town, South Africa, using Raspberry Pis. The Raspberry Pis were programmed to run a Python script that initiated the process of recording every time the device was rebooted. The group was divided into 13 subgroups consisting of two members each, and the map of Intaka also divided into 13 regions where each subgroup was assigned a region to record. On field, the recording devices were placed in areas where the birds vocalized. For every location the device was placed, the GPS co-ordinates and time were recorded. The co-ordinates would then be used to plot a map of the Island indicating regions where the vocalizations were more prominent than others.

The audio files were annotated using Sonic Visualizer, with bird vocalizations labelled with a 1 and non-bird vocalizations, 0.

B. Data preprocessing and transformation

The highest frequency in the audio data was about 11000 Hz and so it was downsampled to 4800 Hz using Nyquist theorem. Mel spectrograms of the audios were constructed using librosa, and transformed into Mel Frequency Cepstral Coefficients (MFCC). MFCCs provide a compact representation of the spectral characteristics of an audio signal, and are widely used in acoustic applications [3]. For this project, 19 coefficients were computed as the MFCC, which were shown to capture much of the important features contained in the mel-spectrogram, with the advantage of having a lower size. For example, the dimensions of the mel-spectrograms were 128 x 345, while the MFCCs was 19 x 345. Figure 1 shows a plot of a mel-spectrogram and its corresponding MFCCs.

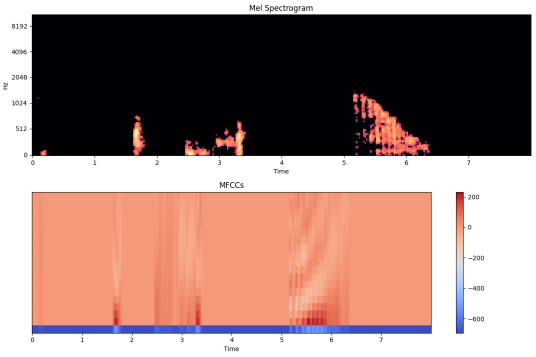

Fig. 1. Visualizations of a Mel-spectrogram (top) and MFCCs (bottom).

C. Data augmentation

As was pointed out in the introduction section, deep learning requires massive amounts of data. Obtaining massive amounts of labeled audio signals, and then transforming them into visual representations proves to be a difficult task. Our data set consisted of just 9092 samples for both training and validation. Fortunately, there is a cheap method that allows us to get around this problem. By applying minor perturbations or

transformations to the original audio signals we can generate new observations. There are several ways to achieve this, but here, we used frequency masking. These masks were applied to random regions of the spectrograms. A total of 8000 augmented samples were added, thus roughly doubling the dataset size. Frequency masks allow the model to learn different representations of the same class, and therefore make it invariant to variations in specific frequency regions.

D. Convolutional Neural Network architecture

The CNN consisted of three convolutional layers with filter sizes of 3 x 3 each. The sizes of the filters and number of layers in the network are constrained to small values since the input has one of the dimensions to only be 19. Choosing higher values would make convolution impossible especially in the deeper layers. For example having a filter size of 5x5 in a fifth layer will not be possible as the input size would be smaller by then.

Each convolution layer was followed by BatchNormalization and MaxPooling layer, each of size.

III. Results

	Predicted	
	Negative	Positive
Actual Negative	125	376
Positive	106	1613

TABLE I
CONFUSION MATRIX

Accuracy	Precision
78.3%	81.1%

TABLE II
MODEL EVALUATION

IV. Discussions

Since the aim of this project is to be able to detect bird vocalizations, the most important metric for evaluating the model is its precision. The test dataset consisted of 2220 observations. From the confusion matrix presented in Table I, out of a total of 1719 vocalization observations, the model correctly predicted that 1613 were actually bird vocalizations. This gives the model a precision of 81.1%.

References

[1] N. Loris, M. Gianluca, B. Sheryl, and P. Michelangelo, "An ensemble of convolutional neural networks for audio classification," *Applied Science*, 2021.

[2] J. Shonfield and E. M. Bayne, "Autonomous recording units in avian ecological research: current use and future applications," *Avian Conservation and Ecology*, vol. 12, no. 1, p. 2, 2017.

[3] A. ZRAR, KH. and A.-T. ABDULBASIT, K., "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, 2022.

.