Bird sound classification with deep learning

Clément Adandé
AI for Science Student
African Institute for Mathematical Science (AIMS)
Muizenberg, South Africa
clementa@aims.ac.za

***Abstract*—Due to factors such as invasion and climate change, animal species, including birds, decline. Then it is urgent to propose solutions to monitoring wildlife. Birds communicate using vocalization and according the place world, bird species is different. This poses a challenge to identify the bird through its vocalization. This study investigate classification algorithms to identify bird sounds. Based on data collected from Intaka Island, two deep learning algorithms are performed : self CNN and transfer learning. We found that using transfer learning is more efficient than convolutional neural networks (CNN).**

***Index Terms*—Deep learning, ecology, CNN, transfer learning, bird sounds**

I. Introduction

The number of many animal species decreases more and more due to species invasion and climate change. This is also observed with birds. Birds communicate by vocalization. Identifying each bird is a very big challenge due to the various birds and contribute to monitoring wildlife. Artificial intelligence can help classify birds, since it continues to gain place in our realities and also in scientific domains. This report relates one aspect of bird sounds classification problem.

The remainder of this document is organized as follows. Section II presents the material used, section III relates literature reviews of bird sound classification, section IV describes the deep learning algorithms used in this project and section V compares these two algorithms.

II. Material

A. Data collection

The classification models performed in this report shared the same dataset. This dataset is built using a bunch of bird sounds recorded in Intaka Island, a harmonious sanctuary for Nature and Urban living, which offers a nest to birds and proves to be a good place for recording bird sounds. Intaka Island is located at Cape Town in South Africa. During a trip make for bird sound classification purpose, each AI for Science's students of AIMS recorded in a time range of 08:00 to 11:00 am, on February third 2024. A collective dataset is created using these recording and is available on Google Drive. Another dataset is built with my own collected sounds and can be found at this Github repository. What are the recording material ?

The recording was succeed using a kit Raspeberry PI (see fig. 1) equipped within a SC card (32 Go). This kit is synchronized with an AIMS server. The recording sounds are with 44100Hz sample rates. To make the easier treatment of data we downsample it.

Fig. 1. Kit Rasberry Pi

B. Data pre-processing

Every recording is downsampled at $22050Hz$ on a frequency range between $1000Hz$ and $9000Hz$ and is used together to create the set of features X and the set of targets Y. Thanks to Sonic Vizualiser, annotations[1] is realised by hand and every label represents either absence (labeled by 0) or presence (labeled 1) according to absence of bird sound or no in a given feature. Each feature corresponds to a normalization of a mel spectrogram made with a Hann analysis window size of 1024 samples, a hop size of 256 samples and mel frequency bins 128 samples. Each feature corresponds to 4 secondes of songs using a sliding window with an overlap of 1 seconde. In other words, each $X[i]$ represents the spectrogram of a sound whose duration is 4 secondes and $Y[i]$ is the corresponding target (0 or 1). The dataset has 1361 examples obtained from the dataset using data augmentation techniques. Furthermore, the size of each spectrogram is 128×345. But why do we use spectrogram instead of audio sounds directly ?

In this report, we trained a CNN on our dataset to learn bird sounds. The expected inputs of a CNN are images. As originally our dataset consists of sounds, we need to convert them into images before feeding them to the model. One way of doing so is by considering the spectrogram

[1]It consists of telling wheather there is bird song in the file or no for computational use.

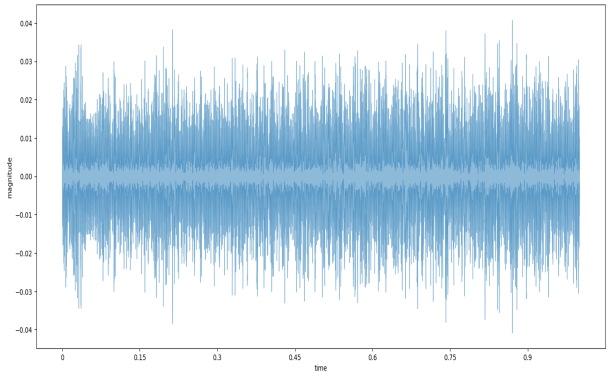

Fig. 2. Wave form of an audio

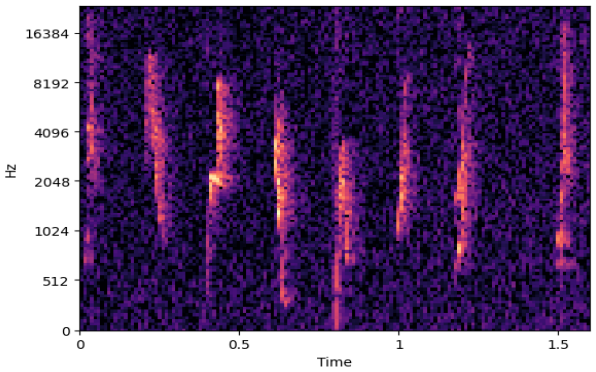

Fig. 3. Spectrogram

associated with each sound. In another words, to reflect what human can hear with our model, we convert audio sound from wave form to spectrogram (see fig. 2 and 3). The transformation of audio into a spectrogram is made using Mel-Frequency Ceptral coefficient [4]–[6]. It's a common technique used in various speech processing techniques. It provides our models sound information which looks like what human could perceives [4].

III. Related literature

Previous work about bird sound classification is described in [1], [2], [7]. In [1], the authors classify 22 bird species using CNN. In [2], the authors use a fused classification method based on the Error Correction Output Coding (ECOC) and Support Vector Machines (SVM) is investigated to classify 11 species of birds.

IV. Methods

In this section, two approaches are proposed to the bird sound classification problem.

A. Self CNN

CNN is usually composed of convolution layer and pooling layer. A convolution layer applies a sequence of filters to translate over the input and pooling layer reduces the dimension of the a subsequent layers. Here is a short description of layers of our model:

- one convolution layer that transports input throught 128 layers using a 4×4 kernel followed by a max pooling;
- one fully-connected layer with 128 neurons and relu activation function;
- one fully-connected layer with 2 neurons and softmax activation function.

As it's a binary classification problem, we used categorical cross entropy loss. For optimisation of the model parameters, we used Adam optimizer.

B. transfer learning

transfer learning is a machine learning (ML) technique that uses a model pre-trained to perform another task. It adapts the learned representations or embeddings to a new data. In the following steps, we describes how it's used here.

1) We use ResNet50V2, a model trained on ImageNet[2] for image classification problem.
2) We remove its output layer.
3) We freeze its weights to avoid their update during the training.
4) We add one layer with 2 neurons and softmax activation function.
5) We train the model.

V. Results

Table I presents the performance of the two deep learning approaches for bird sound classification on the held-out test set. The Self CNN achieved a training accuracy of 93.11% but showed lower generalization with a test accuracy of 70.38%, indicating potential overfitting to the training data. In contrast, the transfer learning approach outperformed the Self CNN with a training accuracy of 94.67% and a test accuracy of 76.75%.

Algorithm	Epochs	Train Accuracy (%)	Test accuracy (%)
CNN	10	93.11	70.38
transfer learning	5	94.67	76.75

TABLE I
Accuracy of test set of different models

This result demonstrates the ability of transfer learning to adapt pre-trained features from a large-scale dataset, enabling efficient learning and better generalization with fewer epochs.

VI. Conclusion

This project investigated the classification of bird sounds using two different deep learning techniques: a Self CNN and transfer learning with a ResNet50V2 model. Using data

[2]The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale.

collected from Intaka Island, the results showed that transfer learning performed better than the Self CNN in terms of both accuracy and efficiency. While the Self CNN achieved high training accuracy, its lower test accuracy indicated difficulty in generalizing to new data. On the other hand, the transfer learning model made better use of pre-trained features, leading to improved accuracy with fewer training steps. These results show the value of transfer learning for bird sound classification, especially when the data is small. Future work could involve refining the pre-trained model, using larger datasets, and testing other methods to improve the results further. This study highlights how artificial intelligence can help monitor wildlife by identifying bird sounds accurately and efficiently.

REFERENCES

[1] Jeantet, Lorène and Dufourq, Emmanuel. (2023). Improving deep learning acoustic classifiers with contextual information for wildlife monitoring. Ecological Informatics. 77. 102256. 10.1016/j.ecoinf.2023.102256.

[2] Xue Han, Jianxin Peng, Bird sound classification based on ECOC-SVM, Applied Acoustics, Volume 204, 2023, 109245, ISSN 0003-682X

[3] S. Molau, M. Pitz, R. Schluter and H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrum," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 2001, pp. 73-76 vol.1

[4] Leitner, Boyang Zhang Jared and Samuel Thornton. “Audio Recognition using Mel Spectrograms and Convolution Neural Networks.” (2019).

[5] N. Bulatović and S. Djukanović, "Mel-spectrogram features for acoustic vehicle detection and speed estimation," 2022 26th International Conference on Information Technology (IT), Zabljak, Montenegro, 2022, pp. 1-4, doi: 10.1109/IT54280.2022.9743540.

[6] Khunarsa, P., Lursinsap, C., Raicharoen, T. (2010). Impulsive Environment Sound Detection by Neural Classification of Spectrogram and Mel-Frequency Coefficient Images. In: Zeng, Z., Wang, J. (eds) Advances in Neural Network Research and Applications. Lecture Notes in Electrical Engineering, vol 67. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12990-2_38

[7] Kahl, Stefan, et al. "Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes." CLEF 2019-Conference and Labs of the Evaluation Forum. Vol. 2380. No. 256. CEUR, 2019.