Audio Based Bird Species Classification

AVES

Arpan Datta, Equbal Hossain, Subham Samanta
(arpandatta17@gmail.com, equbal15000@gmail.com, subhamsamanta98@gmail.com)

July 1, 2022

Abstract

The bird species classifier is a system that is equipped with an area AI technology and uses deep learning methods to store and classify bird calls. The system also provides species classification resources to allow automated species detection from observations that can teach a machine how to recognize whether or classify the species. Non-undesirable noises are filtered out and sorted into data sets, where each sound is run via a noise suppression filter model and a separate classification procedure. High pass filter like Mel-frequency Cepstral Coefficient (MFCC) has been used to convert the audio signal into image. In this paper, we propose using convolutional neural networks on the task of automated bird audio detection in real-life environments to recognize the bird species.

1 Introduction

Intro to bird songs:

The birdsong analysis and classification is a very interesting problem to tackle. Birds have many types of voices and the different types have different functions. The most common are song and 'other voices' (e.g. call-type). The song is the "prettier" — melodic type of voice, thanks to which the birds mark their territory and get partners. It is usually much more complex and longer than "call".Call-type voices include contact, enticing and alarm voices. Contact and attracting calls are used to keep birds in a group during flight or foraging, for example in the treetops, alarm ones to alert (e.g. when a predator arrives). Most often these are short and simple voices.
Example:
The song is a simple lively rhythmic verse with a slightly mechanical sound, e.g. "te-ta te-ta te-ta" or three-syllable with a different accent, "te-te-ta te-te-ta te-te-ta".

Intro to the task:

Now we have a clear idea on bird songs and calls.So come to our task now that is the automatic bird species recognition using audio signal processing which can be defined as the problem of identifying the species of a specific bird from its recorded songs.

How our project is helpful in real-life:

Now come to the point how our task is helpful in the real world.Powerful audio signal processing techniques makes it possible to introduce automated methods for the detection of bird vocalizations which will save time of ecologists for similar task. These techniques do not need researchers/ornithologists to actually see or in contact with the bird.

The problems we may face:

Now come to the problems which we may face while working on our task We assume that this Automatic Bird Identification will face many difficulties like-

1.background noise — especially while using data recorded in a city (e.g. city noises, churches, cars)

2.multi-label classification problem — when there are many species singing at the same time

3.different types of bird songs (as described earlier)

4.inter-species variance — there might be a difference in birdsong between the same species living in different regions or countries

5.data set issues — the data can be highly imbalanced due to bigger popularity of one species over another, there is a large number of different species and recordings can have different length, quality of recordings (volume, cleanliness).

so,our ultimate goal is to overcome these difficulties and fulfil our task using neural networks to recognise the bird species based on audio datas.

2 Literature review

Our proposed project is one of the most interesting work over audio data. Many reputed university's students ,teachers ,researchres had done this task before with different datasets and different methodologies.Also a lot of work had happened to be initiated by the various AI challenges, such as BirdCLEF and DCASE. By studying many research papers we came to a decision that CNN-based models are the most common approach in bird call classification as features can be effectively extracted from spectograms and classified as images; though many of them also used CRNN , RNN. so at first we come to the points how our work differs from others

1. We work on our own created unique data.

2. By diving into many research papers we noticed that others mostly used Convolutional

Neural Networks (CNNs) or Recurrent Convolutional Neural Network (RCNNs) where the gap between CNN-based models and shallow, feature-based approaches remained considerably high.
They proposed though many of the recordings were quite noisy the CNNs worked well without any additional noise removal and many teams claimed that noise reduction techniques did not help.some teams also successfully approached it with semi-supervised learning methods (pseudo-labeling) and some increased AUC by model ensemble.Besides they also had done the works on the features like mel-band energy levels,harmonic contents of the audio.

After studing and analyzing their methods and results we came to a decision that -
We will build our CNN model based on the input features Mel-spectograms(Mel-Spectrogram is computed by applying a Fourier transform to analyze the frequency content of a signal and to convert it to the mel-scale)as by analyzing other research papers we came to a decision that spectogram was the best repesentative of a audio data so to upgrade it we used the spectograms in mel-scale as this modified spectrogram simply ignores the sounds humans do not hear and plot the most important parts .

3. We also modified the noises by creating images out of 10s lasting audios(and it increased final model accuracy by 10%) because the longer the length of the audio from which a spectrogram is created, the more information you get on an image but also the more overfitting your model can become. If your data has a lot of noise or silence, there is a chance that 5 seconds lasting audios will not catch the needed information.

4.And atlast we apply different achitectures of CNN with data augumentation,batch normalization,data normalization,optimizers to find the best model fit and compared the accuracies to come to a decision.

so atlast in short if we want to discuss our project path that will be- at first we will extract the signal processing features and turn them into mel-spectrograms then we will apply CNN architecture on melspectrograms and find the output result that how good the bird species are classified

3 Proposed methodology

Data Collection:

Data collection has been done from here. We have have collected 20 species of bird and each species has approx 200 chirping. Our data consists of over 4000 items and as every species has same number of chirping, so one can say that the data is not class imbalance. As the data has a lot of noise or silence, there is a chance that small duration lasting audios will not catch the needed information. Therefore it was decided to create images out of 10s lasting audios. If more lasting audio is taken then it leads to overfitting. Since the

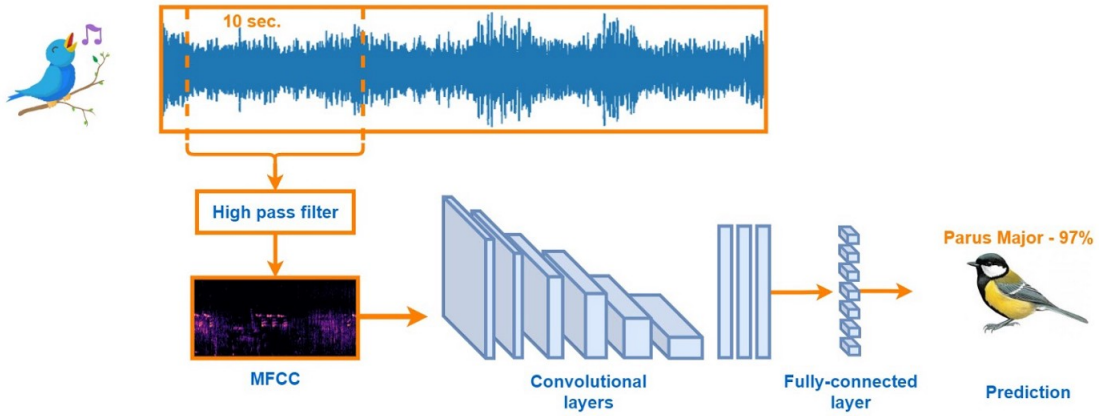


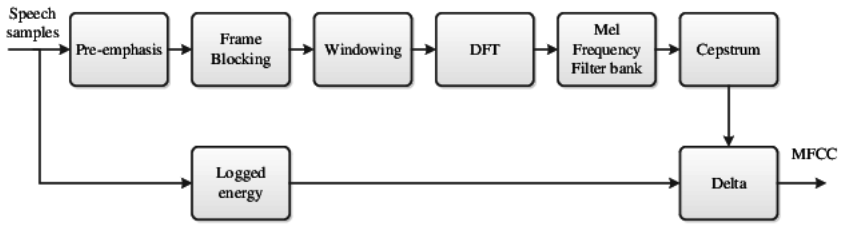
Figure 1: Project Workflow

birds sing in high frequencies, high pass filter was applied to remove useless noise.

We have converted mp3 downloaded data to .wav format to make easy to build the model. The converting process has been done through 'Factory Format' which is itself a model and available in google.

Feature Extraction(using MFCC):

Mel-frequency cepstral coefficients is the full form of MFCC. We have used to convert audio file to image file. The MFCC feature extraction technique basically includes windowing the signal, applying the Discrete Fourier Transformation (DFT), taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse Disrete Cosine Transformation (DCT). The MFCC gives a discrete cosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequency scale.

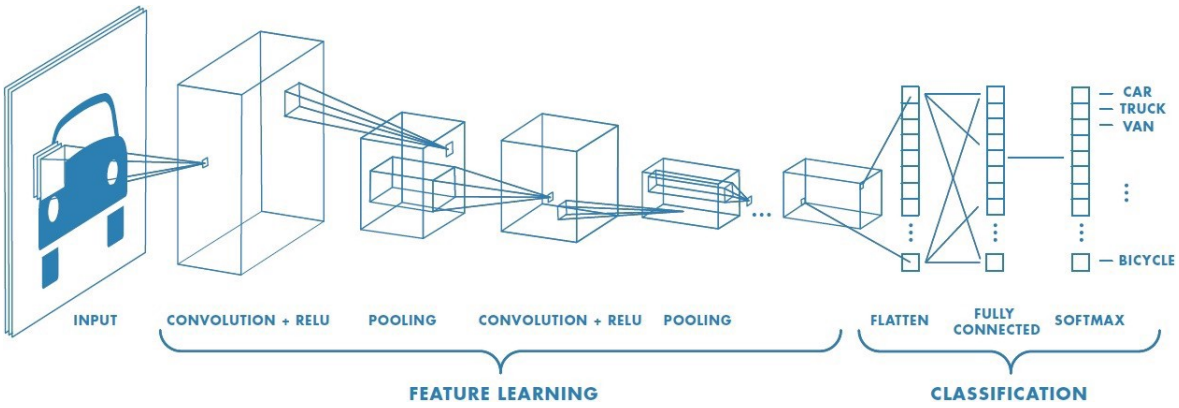


Train-test Split:

Spectro-temporal features (spectrogram) are extracted from the raw audio recordings to be used as the sound representation. After making spectrogram We have taken 70% as training data and 10% as validation and 20% as test data.

Model Fit(CNN):

We use convolution neural network(CNN) as a model with EffeicientNet, ResNet, LeNet, VGGNet and Inception architecture. CNN consists of convolution layer which reduces the size into 224*224*3. Then flatten has been done. Overfitting may occur, so we use dropout to drop some neurons. Basically dropout layer is a mask that nullifies the contribution of some neurons towards the next layer and leaves unmodified all others. In our model we drop 50% neurons of flatten neuron. Then we pass the remaining neuron through fully connected layer(Dense). Dense function is basically a fully connected layer. Finally the neurons pass through softmax activation function and we use Adam as optimizer.



4 Experimental result

- Datasets we have used for our project:
  From the website https://xeno-canto.org we have manually downloaded 200 audio files for each of 20 different species i.e. total 4000 mp3 files. Then we converted those into MFCC images.

- Experimental settings
  At first to fit the model on our input features we use efficientnetB3 CNN model where after using sgd optimizer we get train accuracy 95% and test accuracy 66% after trained the model for 30 epochs whiich clearly denote the overfit. Next by applying batch normalization in same architechture for same number of epochs surprisingly

we get less training acuuracy around 62%.Atlast by applying adam optimizer with data augumentation but no batch normalization and adding 2 hidden layers in fully connected layer we get the training accuracy around 67%. so,finally we apply the above proposed CNN architecture where we use data augumentation with width shift and height shift,a fully connected layer(above described) to find the best accuracies among them.

- Experimental results and comparison with the state-of-the-art methods

For 15-16 species the precision score and recall are greater than 0.60. We can see that the precision score for the species Eurasian Skylark is 1.00 that is no false positive. For Red Shank, Rook Corvas, and Song Sparrow we got 0.97 recall score that means very less false negetive. Except two or three we got average to good precision and recall score and so f1 score.

$$Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Positives}$$

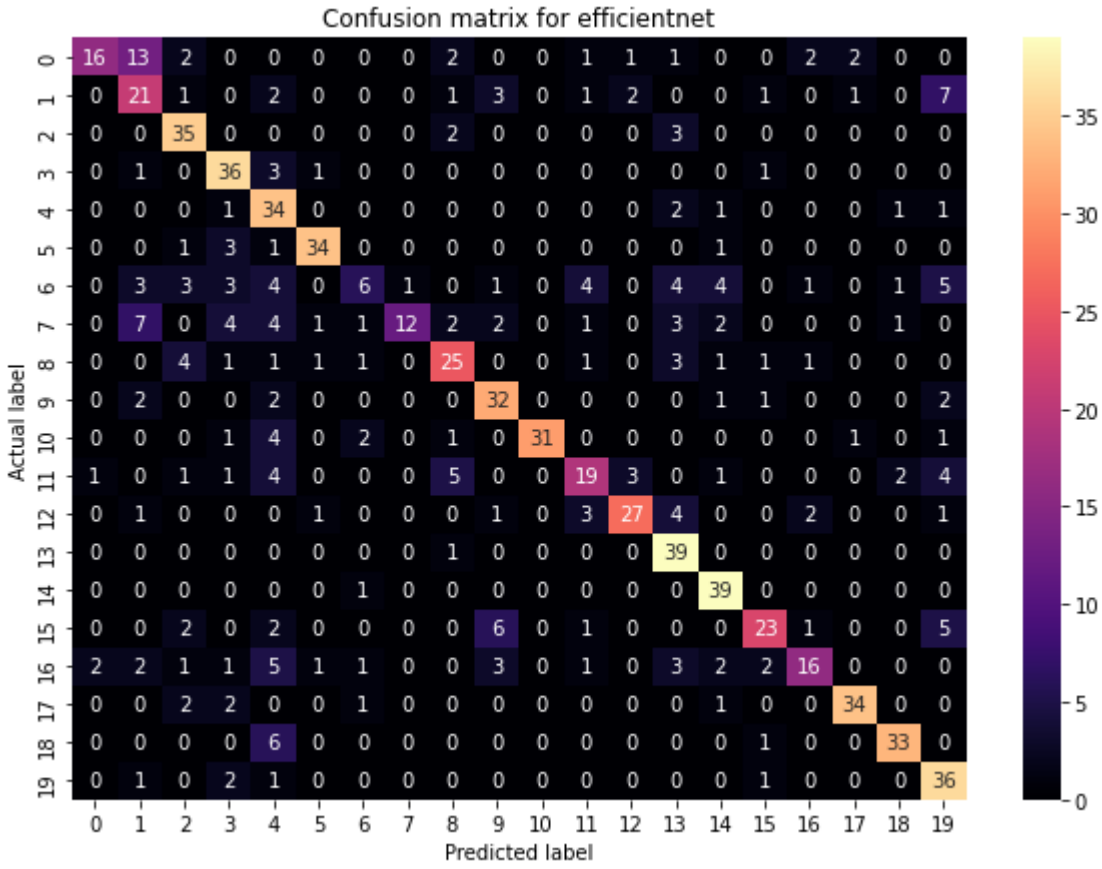
$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}$$

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

	precision	recall	f1-score	support
American_Robbin	0.84	0.40	0.54	40
Blackbird	0.41	0.53	0.46	40
Brown-crested_Flycatcher	0.67	0.88	0.76	40
Chiffchaff	0.65	0.86	0.74	42
Common_Linnet	0.47	0.85	0.60	40
Common_Nightingale	0.87	0.85	0.86	40
Common_wood_pigeon	0.46	0.15	0.23	40
Common_Cuckoo	0.92	0.30	0.45	40
Common_Moorhen	0.64	0.62	0.63	40
Eurasian_Wren	0.67	0.80	0.73	40
Eurosian_Skylark	1.00	0.76	0.86	41
FieldFare Turdus	0.59	0.46	0.52	41
Mistle_Thrush	0.82	0.68	0.74	40
Redshank	0.63	0.97	0.76	40
Rook_corvas	0.74	0.97	0.84	40
Song_Sparrow	0.74	0.57	0.65	40
Song_Thrush	0.70	0.40	0.51	40
Sandpiper	0.89	0.85	0.87	40
Whitethroat	0.87	0.82	0.85	40
YellowHammer	0.58	0.88	0.70	41
accuracy			0.68	805
macro avg	0.71	0.68	0.67	805
weighted avg	0.71	0.68	0.67	805

Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model ("classifier") on a set of test data for which the true values are known. In our classification, We used 800 spectrogram as test data. We got 548 audio as correctly classified and rest of test data have been misclassified.

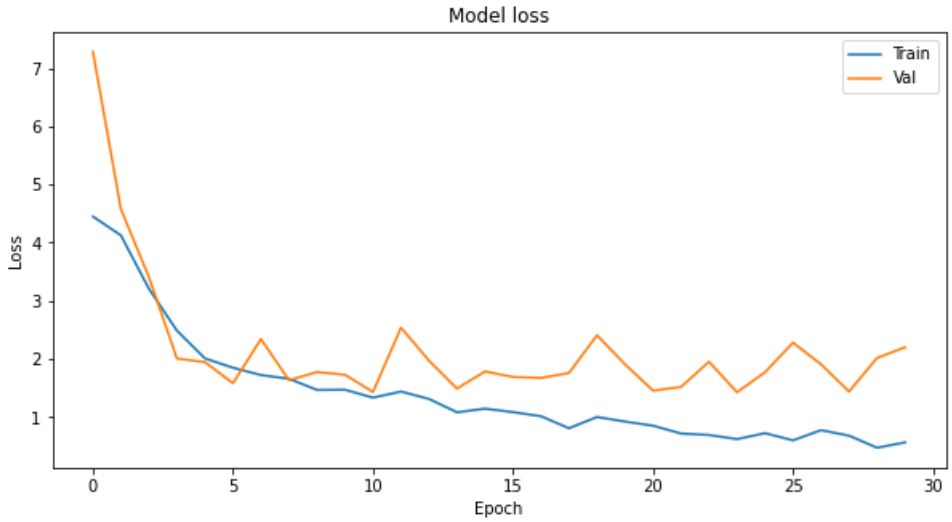


We have used five architecture in our classification problem. The corresponding results given below in Table 1. EfficientNet gives the best train and test accuracy such as 86% and 68% respectively.

Architecture	**Train Accuracy**	**Test Accuracy**
EfficientNet	86.01%	68.07%
ResNet	33.19%	32.29%
Inception	22.15%	23.97%
LeNet	5.89%	5.86%
VGGNet	5.25%	5.22%

Table 1: Train test accuracy

Below plot denotes train and validation loss of efficientnet with respect to epoch. As one can observe with the increasing of epoch, training loss decreasing faster than vali-



dation.

Data Description	**Train Accuracy**	**Test Accuracy**	**Paper Link**
Species(20)*audio(200)	86.01%	68.07%	-
Species(30)*audio(101)	98%	65.5%	here
Species(999)*audio(330)	Not Given	58%	here

Table 2: Comparison Table

And finally we compared our resusts with other research papers and get the above table.
So, after comparing the results one can see that test accuracy is increased in our cases and train accuracy is lesser than others. But it is noted that we worked on different dataset which is unique, so comparing accuracy and taking decision from comparing is not compatible.

- Time Complexity
  O(n)
  In a CNN, the number of features in each feature map is at most a constant times the number of input pixels n (typically the constant is ¡ 1). Convolving a fixed size filter across an image with n pixels takes O(n) time, since each output is just the sum product between k pixels in the image, and k weights in the filter, and k doesn't

vary with n. Similarly, any max or avg pooling operation doesn't take more than linear time in the input size. Therefore, the overall runtime is still linear.

5 Summary

In this project of bird species classification we applied Deep Learning technique of different convolution neural network architectures where input features are the mel-spectrograms collected from audio datasets. By analyzing all the results we came to a decision that efficientnet architecture gives maximum test accuracy.
Achieved Traning accracy(Efficientnet) 86%
Achieved test accuracy(Efficientnet) 68%

In general, the EfficientNet models achieve both higher accuracy and better efficiency over existing other popular CNN architectures, reducing parameter size and FLOPS by an order of magnitude. The other architecture we applied are Lenet , vggnet , Resnet, Inception. But we did not get noticable results from them as their test accuracies are:
Lenet-6%
vggnet-6%
Resnet-33%
Inception-24%
To improve the accuracy we tried different methods like
1.batch normalization
2.data augumentation
3.data normalization
4.optimizers etc.
and atlast collect the best result. Moreover our dataset is unique, nobody worked on this dataset(possibly) and also it has no class imbalance.

Future work:

Atlast we come to the point how we can extend our work in future which we cannot perform for limitation of time.
1.We can concatinate other features like energy levels etc. with mel-spectrograms and apply that as input features in CNN
2.We can use many regularizers like f1score,recall etc. in loss function to improve the acuuracy.
3.We can build our own CNN architechture to improve the results.
4.We can increase the complexity in our project like in this task we classify the different species of birds but if we collect the the bird species country wise and classify them country wise then it will be a challenging task to handle because then it will be a binary class problem contained multiclass datas.

6 Comment

We have collected the data in a very good manner. Therefore data collection part is excellent. Another thing is, we have not find a cause of data augmentation and did not build our own architecture.

References

[1] Sharath Adavanne, Konstantinos Drossos, Emre Çakir, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European signal processing conference (EUSIPCO)*, pages 1729–1733. IEEE, 2017.

[2] Agnes Incze, Henrietta-Bernadett Jancsó, Zoltán Szilágyi, Attila Farkas, and Csaba Sulyok. Bird sound recognition using a convolutional neural network. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000295–000300. IEEE, 2018.

[3] Stefan Kahl, Thomas Wilhelm-Stein, Hussein Hussein, Holger Klinck, Danny Kowerko, Marc Ritter, and Maximilian Eibl. Large-scale bird sound classification using convolutional neural networks. In *CLEF (working notes)*, volume 1866. 2017.

[4] Chih-Yuan Koh, Jaw-Yuan Chang, Chiang-Lin Tai, Da-Yo Huang, Han-Hsing Hsieh, and Yi-Wen Liu. Bird sound classification using convolutional neural networks. In *CLEF (Working Notes)*, 2019.

[5] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019.

[6] A Noumida and Rajeev Rajan. Deep learning-based automatic bird species identification from isolated recordings. In *2021 8th International Conference on Smart Computing and Communications (ICSCC)*, pages 252–256. IEEE, 2021.

[7] Julie Wang Shirley Cheng. Detection of bird species through sounds. `http://cs230.stanford.edu/projects_winter_2021/reports/70762359.pdf`, 2021.

[8] Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. Audio based bird species identification using deep learning techniques. Technical report, 2016.

[9] Willem-Pier Vellinga and Robert Planqué. The xeno-canto collection and its relation to sound recognition and classification. In *CLEF (Working Notes)*, 2015.