Birds Voice Classification using ResNet

Tejashri Manohar Mhatre, Srijita Bhattacharjee

Student, Assistant Professor,
Computer Engineering,
Pillai Hoc College of Engineering and Technology, Rasayni, India

***Abstract*— An area of interest in environment is monitoring animal populations to better understand their behavior, biodiversity, and population dynamics. Audible animals can be automatically classified by their sounds, and a specifically applicative environmental designator is the bird, as it responds quickly to changes in its environment. The aim of this study is to improve accuracy of bird species classifier by using deep residual network, which is implemented and used as a baseline. Literature survey is done on the traditional audio recognition techniques by various researchers in field of deep learning and others. This has not only brought the challenges of understanding the issues in every step of recognition as well as new advancements to overcome the limitations of the previous technique and make it effortless on the user end.**

***Keywords*—Accuracy, Deep Residual Network, Recognition.**

I. INTRODUCTION

Classification is a procedure which is related to categorization. It is the process where ideas are recognized, differentiated, and understood. Classification has many applications and speech recognition is one of the application of it. Speech recognition is involving two or more sub-field of techniques of language computations that develops methodologies which enables the recognition. It is also known as Automatic speech recognition. Some speech recognition system require "Training" where an individual speaker set apart vocabulary in to system. At initial, speech identification was commanded by traditional approaches of recognition systems such as HMM with feed-forward ANN. Now a days, many characteristics of speech recognition have been gripped by methods i.e. deep learning LSTM by RNN.

Birds and their voice is very important for human life. Many people consider sound of birds as a sign for starting of the Eastertide. Many people predict weather and surrounding nature by hearing their sound. Watching bird is also hobby for some people. There are some common species which can be recognize by experts. A goal of this project is to improve accuracy of bird species classifier by using deep residual network, which is implemented and used as a baseline and to design approach for the system that could recognize species of birds automatically.

One challenge that to identify actual sound of bird was recognizing all calls made by a particular specie. A bird's sound which include a remarkable volume of information which contains specie of specific bird and its individual identity, its regional and generative standing, and its possibility of responding its generic reaction of behavior to a prospective receiver. Number of calls and sounds in its range. It was the greatest challenges because of interpreting how exactly birds fluctuate their calls was a notable barricade to interpret how to classify species.

II. LITERATURE REVIEW

Literature survey is done on traditional voice recognition systems.

John Martinsson, [1] set out to upgrade the classification accuracy of the modern bird species classifier. Analyzed methods used are DRNN, multiple-width frequency-delta data augmentation, and fusion of elevation meta-data into the model. By using an analysis of the data set it can be found that the respective number of training samples for each bird species is quite uneven over prediction, from the model of bird species with the most recordings, and that some bird species are harder to classify than others.

Wu, H., Wang, Y., & Huang, J. [2] proposed an algorithm to pick out re-enacted speech. Its Analysis is based on envisage of large extent datasets which shows scattering of the auditory features from human and re-enacted speech is independent. Its method outshout the state-of-the-art methods.

Bottou, L., Curtis, F. E., & Nocedal, J., [3] by using gradient methods it shows its work on topic as a methods of reduction of noice. This work consist of numerical optimization which gives a analysis on the past, present, and future of numeral optimization algorithms in the edge of machine learning applications.

K. Uma Rani, Mallikarjun S. Holi, [4] proposed ARS for bird species which includes partition of feature generation, classifier design and its calculation stages Experiments are based on parametric depiction of syllables using parameter. Its shows appearance related to the density band and content of the sound provide good differentiation capacity within these sounds.

III. DATASET

The BirdCLEF 2014, BirdCLEF 2015 training data is used which is construct from the Xeno-Canto collaborative database. It contains sound recordings of bird species with its normalization frequency sample at 44.1 kHZ, 16 bit .wav format. Training data is supplemented using XML-files which containing metadata. This metadata includes audio recorder ID, Bird species name, Family, order, vernacular name etc. workflow consist of four main steps. First, it needs to preprocess audio files in such a format which can be used to train neural network. In this step audio files get normalized. Secondly, it needs to separate birds song recordings in to different classes i.e. signal and noise. The separation lets us train the neural network on the most relevant data, and it gives us access to a noise class which can be used to augment training samples. After that each audio files get split in to 3-Seconds segments. The noise segments can later be used to augment the training samples shown to improve which should improve generalization.

IV. PROPOSED SYSTEM

This system, aiming at achieving accuracy of individual birds Species detection just by their voice. System propose to use advanced learning techniques to train data i.e. deep Learning. Deep Learning introduces a deep neural network with a maintenance of a Deep Residual Network, which helps to maintain tendency of results. Residual network make it easy for network layers to represent identity mapping. Helps to give qualitative results.

V. WORKFLOW

Its workflow consist of four main steps. First, it needs to preprocess audio files in such a format which can be used to train neural network. In this step audio files get normalized. Secondly, it needs to separate birds song recordings in to different classes i.e. signal and noise. The separation lets us train the neural network on the most relevant data, and it gives us access to a noise class which can be used to augment training samples. After that each audio files get split in to 3-Seconds segments. The noise segments can later be used to augment the training samples shown to improve which should improve generalization.

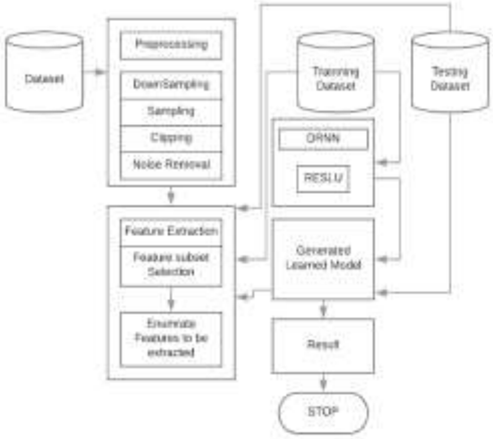

Figure 1. System Architecture

A. Preprocessing

Birds sound recordings first down sampled to 22,050 HZ using the linux command-line tool sox. Which down samples the audio files in directory, after that split down sampled audio files in to signal and noise.

B. Training

Models are trained using modified version of Keras ImageDataGenerator. To train residual neural network it needs to take input from selected dataset. Training raised to calculate the best set of weights for increase neural networks accuracy. The more training the more accuracy will get. Results of this training is framed using four parameter such as, training accuracy, training loss, validation accuracy, validation loss. In figure 2 and Figure 3 it shows the graph of validation and training. Its accuracy is displayed in table i.e. Table I shoes training loss and accuracy and Table II shows validation loss and accuracy.

C. Audio Detection

- The signal part is extracted by first computing signal mask to extract the relevant part of sound wav, where vi=0 indicates that xi is not part of the signal, and vi=1 indicates that it is part of the signal. The mask is derived from binary image which is computed by analyzing the normalized amplitude spectrogram of x Let b_j be the binary image, and let s_i be the normalized range spectrogram of x where b_j and s_i have the same dimensions.
- The pixel at index (i, j) in the binary image is then set to one if s(i) is t times larger j than the row and column median of ?s at row i and column j .
- The binary image is further processed by applying a binary erosion followed by a binary dilation on the image, both using a kernel size of 4 by 4, which smooth out the regions identifiable as bird vocals and the bird vocals mask v¯is derived from the binary image by setting vj to one if the column b ¯j contains one.
- The mask is also smoothed by performing two more binary dilutions (kernel size 4), and is then re-scaled such that |v = |x |.The signal mask is computed by setting the threshold t = 3, and the noise mask is computed by setting t = 2.5 and then inverting the mask at the end (flipping 0s to 1s, and 1s to 0s). This may leave parts of the wave which are identifiable as neither signal nor noise (2.5-3). These parts are inspect to not promote with any admissible information for the network, and they are simply ignored.

D. Data Augmentation

Data augmentation is a way of increasing number of training samples in a dataset by augmenting the training data. Even if the BirdCLEF dataset is one of the largest bird song dataset is available, the number of training samples per bird species is rather small; on average around thirty samples per sound class. Data augmentation can also be used to form the training samples harder by introducing noise to the signal, which is intended to prevent over fitting and make the model discover better due to a higher noise invariance same Class and Noise addition in order to improve the convergence rate of the neural network, the training samples are augmented by actively combining each sample with another same class sample, which lets the network see more relevant data at once. The samples are also add-on with three random noise segments, to build the network more noise invariant, and therefore discover better. Each sample shown to the neural network is thus a combination of two Randomly chosen same class signal segments x and x2 , and three randomly chosen noise segments n, n1 and n2 of the same length. Where α ∈ [0,1] is pick out evenly at random, β = 0.4 is utilize as a dampening factor of the noise segments.

E. Equations

Computations of binary image can be analyze normalized amplitude spectrogram, which define as:

$$b_j^{(i)} = \{1,\ \ if\ \ s_j^{(i)} > t \times median(\overline{s^{(i)}}) \cap s_j^{(i)} > t \times median(\overline{s_j}) \quad , \\ 0, otherwise \qquad (1)$$

Samples are segmented using three randomly chosen values, these sample segments are calculated by using following equation

$$\overline{s_{aug}} = \propto \bar{x_1} + (1-\propto)\bar{x}_2 + \beta(\bar{n}_1 + \bar{n}_2 + \bar{n}_3), \qquad (2)$$

Here it uses cross entropy loss function to measure the performance of neural network model. Following equation specifies the cross entropy function.

$$C = -\frac{1}{N}\sum_{N=1}^{N}[y_n \log y_n + (1 - y_n) + (1 - y_n)], \quad (3)$$

To compute local static information about birds vocals and to find how signals can be process further it uses multiple-width -frequency-delta data augmentation is used. This can be define as:

$$d_t = \frac{\sum_{k=1}^{k} = k(x_{t+k} - x_{t-k})}{2\sum_{k=1}^{k} k^2} \qquad (4)$$

F. Architecture

The architecture of the convolutional neural network uses 18 layer, which includes initial convolution and normal input layer, then eight similar basic blocks with different configurations, where block consists of a convolution 2D, and max pooling layer 2D. Finally, the network has a dense layer and a softmax layer.

- Convolution2D layer uses 64 7x7 kernels, 2x2 stride configuration with (128, 256, 64) output shape.
- Maxpooling2D layer uses 3x3 kernel, 2x2 stride configuration with (64,128, 64) out shape.
- Basic block input layer uses 64 3x3 kernel, 1x1 stride with (64, 128, 64) output shape
- Basic block input layer uses 128 3x3 kernel, 2x2 stride configuration with (32, 64, 128) output shape.
- Basic block input layer uses 256 3x3 kernel, 2x2 stride configuration with (16, 32, 256) output shape.
- Basic block input layer uses 512 3x3 kernel, 1x1 stride configuration with (8, 16, 512) output shape.
- AveragePooling2D uses (8x16) pool size with (1,1, 512) output shape.
- Final layer consist of dense layer and softmax layer.

VI. RESULT ANALYSIS

Result of this system can be display as matching i.e. identification of classes. It performs different concepts for prediction of bird class. As a result it considers identification and predicted output as a result of this model. The accuracy of model diverge for different classes of birds sound. The model has a decent accuracy for ten sound classes on which it gives accuracy of 90%. As out of ten birds classes it identifies nine accurate classes. A perfect accuracy would result in identification of accurate ten bird classes.

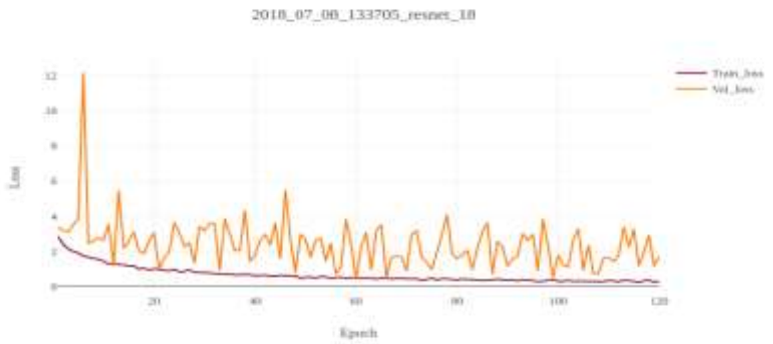

Figure 2. Loss of training and validation.

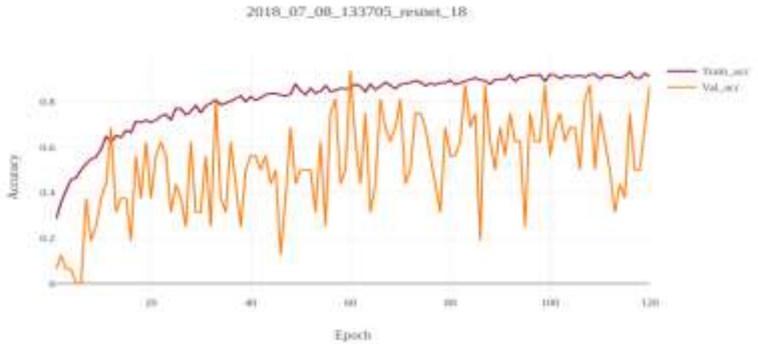

Figure 3. Accuracy of training and validation.

TABLE 1 TRAINING LOSS AND ACCURACY

NO	Training	
	Loss	***Accuracy***
1	0.8360	0.7192
2	0.5350	0.7922
3	0.4183	0.8529
4	0.3964	0.8549
5	0.3648	0.8618
6	0.3838	0.8708

TABLE 2 VALIDATION LOSS AND ACCURACY

NO	Validation	
	Loss	***Accuracy***
1	1.4769	0.5000
2	1.3174	0.6875
3	1.1725	0.8125
4	1.1560	0.6250
5	1.4157	0.5625
6	2.9584	0.5625

VII. CONCLUSION

Residual Neural Network is effective approach for audio classification. Outperforms existing open source implementation Can be an effective alternative to other existing methods Proposed architecture experimentally seemed to be more scalable as with the increase in the number of classes. It was able to comfortably beat the existing open source implementation With the availability of more computing resources, the effectiveness of the presented method can be further improved since, deep learning method is effective but at the same time its computational very expensive.

VIII. ACKNOWLEDGMENT

I would like to acknowledge to all who gave guidance, timely advice and support to complete this work.

REFERENCES

[1] J. Martinsson, “Bird Species Identification using Convolution Neural Networks” Department of Computer Science and Engineering, Chalmers University Of Technology, University Of Gothenburg, Gothenburg, Sweden 2017.

[2] Wu, H., Wang, Y., & Huang, J. “Identification of Reconstructed Speech. ACM Transactions on Multimedia Computing, Communications, and Applications”, Shenzhen University, Sun Yat-Sen University, Guangdong Polytechnic Normal University. 2017.

[3] Bottou, L., Curtis, F. E., & Nocedal, J., “Optimization Methods for Large-Scale Machine Learning.” SIAM Review, 60(2), 223–311. 2018

[4] K. Uma Rani, Mallikarjun S. Holi, “A Comparative Study of Neural Networks and Support Vector Machines for Neurological Disordered Voice Classification”, International Journal of Engineering Research &amp; Technology (IJERT), ISSN: 2278-0181, April – 2014.

[5] P. Forczmanski and T. Maka, “Investigating Combinations of Visual Audio Features and Distance Metrics in the Problem of Audio Classification,” International Conference on Computer Recognition System CORES 2016.

[6] T. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, “Multichannel signal processing with deep neural networks for automatic speech recognition,” IEEE/ACM Trans. Audio Speech, Lang. Process., 2017.

[7] A. Moosavian, H. Ahmadi, A. Tabatabaeefar and M. Khazaee “Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing,” Department of Mechanical Engineering of Agricultural Machinery, University of Tehran, Karaj, Iran Shock,and Vibration, 2013, IOS Press.

[8] V. Elaiya raj a, P. Meenakshi sundaram, “Audio Classification Using Support Vector Machines and Independent Component Analysis,” Journal of Computer Applications, Volume V, Issue 1, 2012.