International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Birds Audio Detection using Convolutional Neural Network and Transfer Learning

Greeshma C Shekar[1], Chandu B L[2]

[1]*Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, Karnataka, India*
[2]*Department of Computer Science and Engineering, AMC Engineering College, Bangalore, Karnataka, India*

Abstract - *Convolutional neural networks (CNNs) are prominent toolkits of machine learning, which have proven to be very efficient in the fields of image and sound recognition. Bird audio detection aims to detect whether there is a bird sound in an audio recording or not. Many birds are preferably detected by their sounds, and thus, passive acoustic monitoring is appropriate. A lot of things can be represented as images, which implies that an image recognizer can learn to complete many tasks, in this case, a sound can be converted to a spectrogram. A spectrogram is basically the visual representation of the signal strength/loudness, of a particular signal over time at different ranges of frequencies present in a specific waveform. In this paper, birds' audio is detected using Convolutional Neural Network and Transfer Learning.*

Key Words*: *Convolutional Neural Networks, Artificial Intelligence, Machine Learning, Deep Learning, Spectrogram, and Transfer Learning.

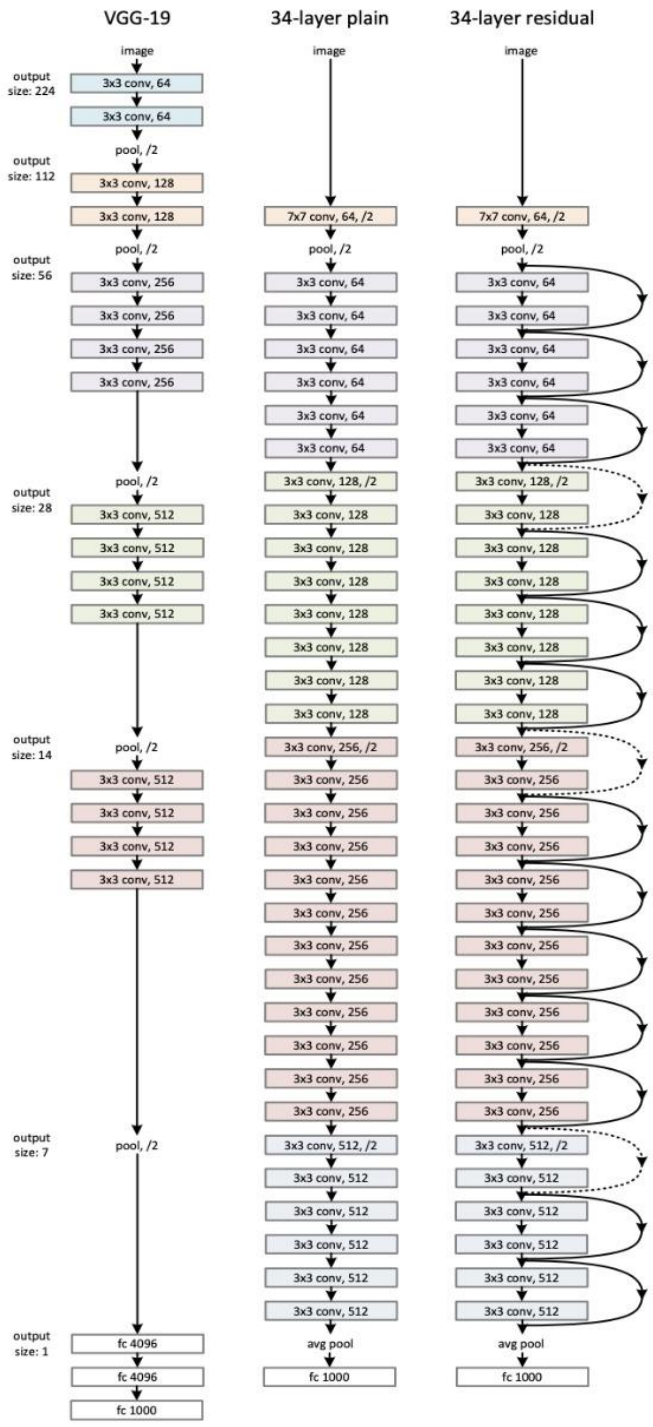
1.INTRODUCTION

In the past few years, algorithmic sound recognition has relished a constant increase in interest. The acceptance of deep learning and the numerous types of neural networks have furnished a new under-explored mechanism of advancing such classification problems.

Detection of bird sounds in audio is one of the most important tasks for automatic wildlife monitoring. The present generation of software tools requires manual work from the user: to choose the algorithm, to set the settings, and to post-process the results.

An increasing effort over the past few years for monitoring vocalizing species has been a valid indicator of biodiversity. Monitoring avian populations in their habitats is one of such efforts since birds are good ecological indicators of environmental changes.

Resnet34 is basically a 34-layer convolutional neural network that can be used as a state-of-the-art image classification model. The Resnet models we use have been pre-trained on the ImageNet dataset, which contains 100,000 images across 200 classes.



We use a process called Transfer Learning, in which we start our model from the pre-trained checkpoint and fine-tune our Resnet model from this base state.

We made use of fastai, a free open-source library for deep learning. This library is built on top of PyTorch, and it provides APIs to the important deep learning applications as well as data types.

This paper is organized as follows:

1.1 Related Work

Provides information about the concerning related work with required data.

1.2 Methodology

This section, describes data and methods for bird sound detection using the graphs for better understanding of the methodology.

1.3 Result and Discussion

Elucidates the experimental results and discussions of the model.

1.4 Conclusion

This section concludes the paper finally.

2. RELATED WORK

CNN can act as a feature extractor which is proven to be superior in many classification tasks. The visual representation of audio in the form of the spectrogram is very popular as an input feature as it contains many channels of information such as channel, environment, and so on. State-of-the-art results can be obtained only if the hyperparameters of the CNNs are carefully tuned [1].

Deep CNN architectures perform better as compared to the standard CNN model [2,3]. In the past variety of CNN architectures have been tested for bird audio detection. We make use of the pre-trained Resnet34 architecture to easily build a classifier in fastai which achieves an accuracy of 88.43%.

3. METHODOLOGY

3.1 Datasets - We used the dataset which comes from a UK bird-sound crowdsourcing research spinout called Warblr. From this initiative, we found 10,000 ten-second smartphone audio recordings from around the UK. The total audio duration is about 44 hours. The audio covers a wide distribution of UK locations and environments. Warblr automatically recognizes British birds by their song.

	itemid	hasbird
0	759808e5-f824-401e-9058	1
1	1d94fc4a-1c63-4da0-9cac	1
2	bb0099ce-3073-4613-8557	1
3	c4c67e81-9aa8-4af4-8eb7	1
4	ab322d4b-da69-4b06-a065	0

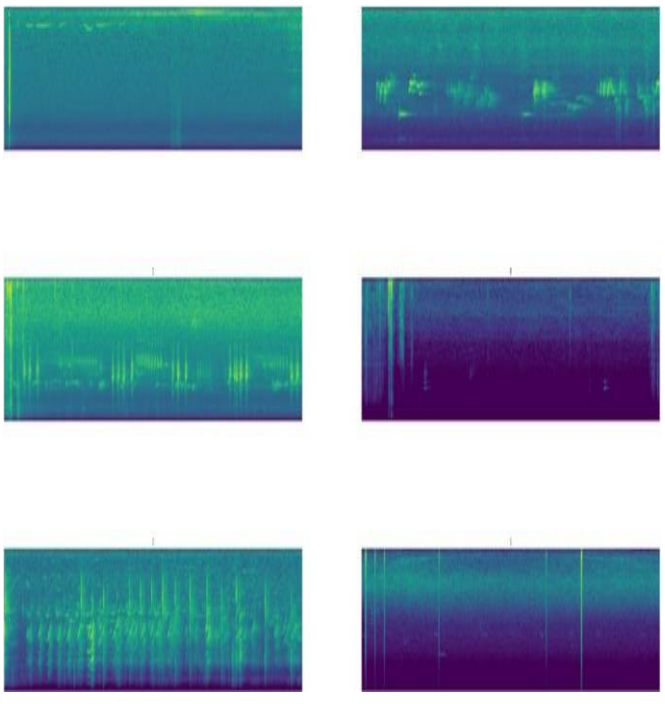
3.2 Methodology - We take the 10-second audio clip and obtain the MelSpectrogram, which depicts the acoustic time-frequency representation of sound. The power spectral density P (f, t), is sampled into several points around equally spaced times t_i and frequencies f_j on a Mel frequency scale.

The Mel frequency scale is defined as follows:

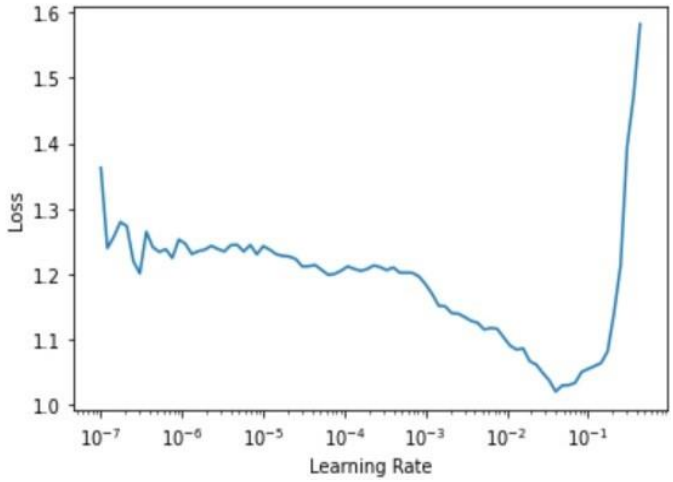
Mel = 2595 * log10 (1 + hertz / 700)

and the inverse of it is:

hertz = 700 * ($10.0^{mel / 2595.0}$ - 1).



We resize each of the spectrograms to the size (120, 800) by cropping them using the DataBlock API of fastai to quickly build a DataLoader and assign a label of 1 or 0 according to the presence of the birds singing sound as provided by the CSV file provided along with the dataset. We normalize the spectrogram images using batch transformation by making use of the GPU.

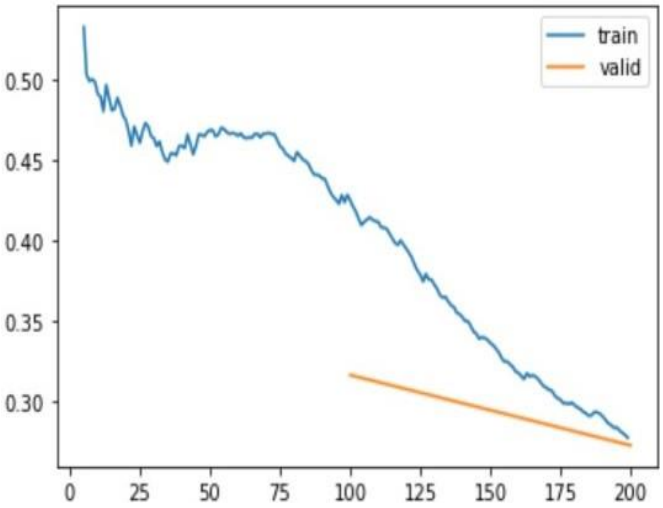


The dataset is randomly split into training and validation set by 80% and 20% respectively. We make use of the pre-trained ResNet34 weights which is fine tuned as best as possible. We make use of the Learning rate finder to make sure we have the right learning rate, whose basic idea is to start with a very small learning rate for one mini batch and find what the loses are later on, then we double it each time and use another mini batch and make note of the loss until the loss gets worse. It is advisable to take the last point where the loss was clearly decreasing. This brilliant idea was proposed by the researcher Leslie Smith.

We fine-tune our model with the learning rate which we just found using the learning rate finder. We know that the Convolutional Neural Network consists of many layers, of which the final layer uses a matrix with enough columns such that the output size is the same as the number of classes in our model.

The final layer is specifically designed to classify the categories in an original pre-trained dataset. This final layer is of no use for us, so when we do the transfer learning, we remove it and replace it with a linear layer that has two activations.

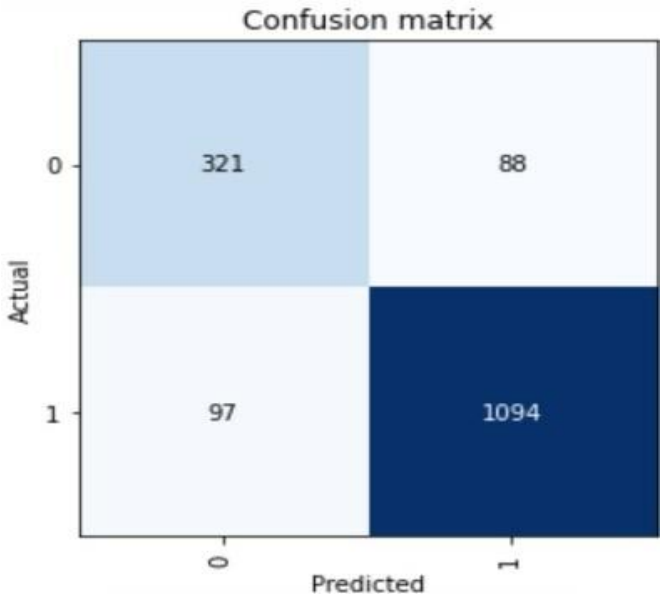
This newly added layer contains random weights but the other layers have been carefully trained for image classification, so it trains the randomly added layers for one epoch with all other layers frozen, then unfreezes all of the layers for two epochs.



As you can see the training loss keeps getting better and better and eventually the validation loss slows down, if we continue further it may start to overfit so, we stop the training after two epochs but, what matters, in the end, is the accuracy or the chosen matric and not the loss.

4. RESULTS AND DISCUSSION

Now let's see the mistakes the model is making creating a confusion matrix. The rows represent the presence and absence of the bird sounds, while the columns represent the predicted output of the model and the diagonal of the matric depicts the outputs which were correctly classified. The off-diagonal cells represent those that were classified incorrectly.



Our proposed model obtained an accuracy of 88.43% and we assume that the model's performance can be improvised further by including more non-bird audios in the dataset.

5. Conclusion and Future Scope

By using Transfer Learning, it is much easier to build state-of-the-art Deep Learning models without the need for many GPUs and large datasets. We can make use of libraries like fastai to quickly build useful models and deploy them in real life.

The future scope includes:

[1] We propose that the model's performance can be improvised further by including more non-bird audios in the dataset.

[2] The creation of an android application will practically be more useful to the users.

6. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097– 1105, 2012.R. Solanki, "*Principle of Data Mining*", McGraw-Hill Publication, **India**, pp. **386-398, 1998.**

[2] K. He et al., "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, No. **770-778, 2016.**

[3] G. Huang et al., "Densely connected convolutional networks," in Proceedings of Computer Vision and Pattern Recognition, **2017**, pp. **4700–4708.**

[4] M. Lasseck, "Towards automatic large-scale identification of birds in audio recordings," in Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction - **6**th International Conference of the CLEF Association, CLEF , **2015**, pp. **364–375**.

[5] E. Cakir et al., "Convolutional recurrent neural networks for bird audio detection," in Proceedings of the European Signal Processing Conference (EUSIPCO), **2107**, pp. **1794–1798**.

[6] C. Szegedy and S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:**1502.03167, 2015**.

[7] I. Potamitis, "Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity," Ecological Informatics, vol. **26,** pp. **6–17, 2015.**

[8] Y. Bengio and X. Glorot, "Understanding the difficulty of training deep feedforward neural networks," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. **249–256**, **2010**.

[9] A. Torfi et al., "3D convolutional neural networks for cross audio-visual matching recognition," IEEE Access, vol. **5**, pp. **22 081–22 091, 2017.**

[10] I. Teivas, "Video event classification using 3D convolutional neural networks," Master's thesis, Tampere University of Technology, **2016.**

AUTHORS PROFILE

Miss. Greeshma C Shekar is pursuing Bachelor of Technology degree in department of Computer Science and Engineering from Dayananda Sagar University, located in Bangalore, Karnataka, India. She has published a paper entitled "Blockchain and Cryptocurrency: The world of Blockchain and Cryptocurrency" in IJCSE Volume-9 Issue-7, Jul 2021. She is currently in her final year of Engineering and will be graduating from Dayananda Sagar University in the year 2022.

Mr. Chandu B L *is* pursuing a Bachelor of Engineering in department of Computer Science and Engineering from AMC Engineering College located in Bangalore, Karnataka, India. He is currently in his final year of Engineering and will be graduating in the year 2022.