Bird Species Classification based on Images and Audios using Deep Learning

Jahnavi S, Keerthana K, Priyanka S, Sanjana L, Dr. Shilpa B L

Dept. of Information Science and Engineering (of VTU Affiliation), GSSS Institute of Engineering and Technology for Women, Mysuru, India.

Abstract **– Birds play a crucial part in our ecosystem. Accurate bird classification is essential for understanding ecosystem dynamics, assessing biodiversity, and guiding conservation strategies. Traditional methods relying on manual observation are time-consuming and limited in scope, hindering comprehensive ecological studies and conservation efforts. In this paper, we propose a Convolutional Neural Network (CNN) approach for the automated classification of bird species from images or audio. The visual component of our model uses VGG16 (a specific CNN architecture) to extract features from bird images, while the Multilayer Perceptron (MLP) classifier extracts audio features. Our methodology leverages the power of deep learning to recognize distinct features of bird species, thus enabling accurate identification. We employ a dataset comprising images of various bird species. The dataset is pre-processed to eliminate noise and enhance image and audio quality, thereby facilitating effective training of the CNN model. The architecture is designed to extract hierarchical features from input images and audio and learn discriminative patterns associated with different bird species. Experimental results demonstrate the effectiveness of our proposed approach in accurately identifying the bird species. The model's performance is evaluated using standard metrics such as accuracy. The proposed model promises potential for applications in ecosystem monitoring and environmental conservation.**

Keywords- bird species, species detection, deep learning, Convolutional Neural Network (CNN), VGG16, image classification, audio classification, ecosystem monitoring, biodiversity conservation.

1. INTRODUCTION

As integral components of ecosystems, birds play multifaceted roles, including seed dispersal, pollination, and insect control, thereby exerting profound influences on ecosystem dynamics and functioning. Furthermore, birds serve as indicators of environmental health, with changes in bird populations often reflecting broader ecological shifts. Understanding bird species distributions, behaviours, and interactions is essential for analysing conservation strategies and safeguarding biodiversity. Traditional methods of bird classification predominantly rely on manual observation, which is labor- intensive, time-consuming, and prone to human error. In our model, we address the challenges of limited labelled data by employing data augmentation strategies such as cropping, rotation and flipping, which help to augment the training dataset, thereby reducing overfitting and improving the generalization ability of the model.

Deep learning implemented in our model, has revolutionized various fields, including computer vision and audio processing, by enabling machines to automatically learn hierarchical representations from raw Data. In recent years, Convolutional Neural Networks (CNN) have demonstrated remarkable performance in image classification tasks, while Multilayer Perceptron (MLP) has shown promising results in audio-related tasks such as speech recognition and sound classification.

A. Motivation

The motivation behind this research is driven by the need for efficient and accurate methods to monitor avian populations and their habitats. Accurate identification of bird species is crucial for assessing the population, habitats, and other details of birds. The endeavour to classify bird species through images and audio using deep learning is fueled by the need for effective, scalable methods for monitoring avian populations and ecosystems. Our methodology focuses on designing an efficient CNN architecture capable of capturing fine-grained features of different bird species.

B. Problem Statement

The traditional methods of bird classification rely on manual observation, that are often time-consuming and reliant on expert knowledge, posing limitations in scale and accuracy. These challenges enforce the need for automated approaches for species detection and classification of birds. By leveraging deep learning techniques, the goal is to overcome the limitations of traditional methods.

C. Objectives

- To identify the bird species based on image input or audio file input.
- To implement transfer learning model vgg16 to train the bird's species image dataset and to achieve better accuracy.
- To implement species classification from bids sounds, we implement an MLP classifier for training sound files.
- To implement a Python-based flask web framework a user interface where users can upload bird images or bird sound files and the classification result is displayed with the bird species information.

TABLE I. RELATED WORK

Paper title	***Methodology***	***Dataset Used***	***Key Findings***
Audio Classification of Bird Species Using Convolutional Neural Networks	CNN to identify patterns in the spectrograms specific different bird species	Large collections of labelled bird song recordings	CNNs achieve high accuracy in classifying bird species from audio data.
Eyebirds: A Smartphone App for Water Bird Recognition	Deep learning model (AM-CNN) with an attention mechanism to focus on crucial visual featuresin bird images.	Waterbird image dataset containing over 548 species across 48 families and 203 genera	Eyebirds app effectively identifies water bird species using smartphone photos.
Acoustic Classification of Bird Species Using Early Fusion of Deep Features	Fuse these deep features from various models using early fusion.	A dataset containing recordings of 43 bird species	Early fusion of deep features from pre-trained CNNs improves bird call classification accuracy.
Classifying Bird Species Audio Using Data Augmentation and Convolutional Neural Networks	New bird calls are based on the learned patterns from the CNN.	Existing bird song collection is augmented using data augmentation techniques.	Data augmentation techniques increase the size and diversity of the training data, leading to improved CNN performance.
Leveraging Audio-Visual Cues: A Multi-Modal Approach to Bird Species Classification with CNN and Multiple Kernel Learning	Train a final classifier to predict bird species based on the fused audio-visual features.	A large dataset containing synchronized audio recordings and images of various bird species	Combining audio and visual information improves bird species classification accuracy compared to using only audio or visual data alone.
Multi-Label Bird Species Classification: A Transfer Learning Approach	Utilize a pre-trained deep learning model (e.g., ResNet, Inception) on a large image dataset (e.g., ImageNet).	Custom image dataset containing bird images annotated with multiple bird species labels (e.g., birds appearing together at a feeder).	Transfer learning from pre-trained models significantly reduces training time compared to training from scratch.
Accelerating Bird Species Identification: A Deep Learning Approach on GPUs	CNN learns to identify distinctive visual features (color patterns, beak shapes) that differentiate bird species	Large bird image collections (e.g., Caltech-UCSD Birds-200)	Utilizing GPUs significantly accelerates the training process for deep-learning models in bird species identification.
Machine Learning Classification of Bird Species Based on Vocalizations	Preprocess bird song recordings: convert them into spectrograms (visual representation of sound).	Large collections of labeled bird song recordings (e.g., Xeno- canto)	Machine learning models can effectively classify bird species based on the acoustic features extracted from their songs.
Automated Recognition of Endemic Bird Species Using Deep Learning	Compile a dataset of images containing endemic bird species from the target region.	Custom image collection focusing on endemic bird species of the target region (potentially combined with existing datasets like Birds-of-the- World)	Deep learning models can accurately recognize endemic bird species from images.
PakhiChini: Deep Learning for Automatic Bird Species Identification	Utilize a pre-trained deep learning model (e.g., Inception V3) to extract features from bird images.	Large bird image dataset	Deep learning models effectively learn image features for bird species classification.

This summary provides an overview of the methodologies, datasets and finding of selected papers on CNNs in bird species detection, and offering insights into challenges.

2. MATERIALS AND METHOD

Data Collection and Pre-processing:

Gather a set of bird images sorted by species, maintaining uniformity in image dimensions (e.g., 224x224 pixels), standardizing pixel values, and possibly enhancing the dataset through augmentation techniques to bolster training set variety.

Model Selection:

Choose the VGG16 model as the base for your classification project, valued for its straightforwardness and effectiveness in image classification. Employ either the pre-trained VGG16 model or adapt it to suit your particular dataset through fine- tuning.

Model Architecture:

Develop a fresh classification component to accompany the VGG16 core, typically consisting of one or more fully connected layers followed by a softmax layer to produce class probabilities. Modify the neurons in the final dense layer to align with the number of bird species in your dataset when fine-tuning VGG16.

Training:

Divide your dataset into training, validation, and testing sets, each serving specific roles in model training, hyperparameter tuning, and performance assessment. Train the model using the training set, validate its performance on the validation set, and monitor crucial metrics such as accuracy and loss.

Evaluation:

Evaluate the model's performance after training using the testing set, computing metrics like accuracy to measure its ability to generalize. Examine prediction visualizations and misclassifications to pinpoint model constraints and potential areas for enhancement.

Fine-Tuning:

Consider fine-tuning the model if its performance is below expectations, and modifying hyperparameters such as learning rate, batch size, or regularization strength to improve efficacy.

Deployment:

Once content with its performance, proceed to deploy the model for inference on new bird images. This may entail incorporating it into platforms such as a Tkinter application, allowing users to upload bird images and obtain predictions.

3. PROPOSED METHOD/ ALGORITHM

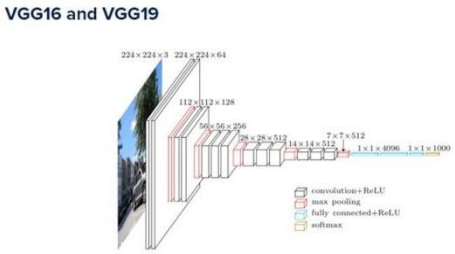

Fig 3.1 **VGG16 architecture**

Each layer in a CNN applies a different set of filters, typically hundreds or thousands of them, and combines the results, feeding the output into the next layer in the network. During training, a CNN automatically learns the values for these filters. In the context of image classification, our CNN may learn to Detect edges from raw pixel data in the first layer. Use these edges to detect shapes (i.e., "blobs") in the second layer. Use these shapes to detect higher-level features such as facial structures, parts of a car, etc. in the highest layers of the network. The last layer in a CNN uses these higher-level features to make predictions regarding the Contents of the image. In terms of deep learning, an (image) convolution is an element-wise multiplication of two matrices followed by a sum.

1. Take two matrices (which both have the same dimensions).

2. Multiply them, element-by-element (i.e., not the dot product, just a simple multiplication).

3. Sum the elements together. Kernels Again, let's think of an image as a big matrix and a kernel as a tiny matrix (at least to the original "big matrix" image), depicted in the above Figure. As the figure demonstrates, we are sliding the kernel (red region) from left to right and top to bottom along the original image. At each (x;y)-coordinate of the original image, we stop and examine the neighborhood of pixels located at the center of the image kernel. We then take this neighborhood of pixels, convolve them with the kernel, and obtain a single output value. The output value is stored in the output image at the same (x;y)-coordinates as

4. the center of the kernel. If this sounds confusing, no worries, we'll be reviewing an example in the next section. But before we dive into an example, let's take a look at what a kernel looks like:

$$K = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Figure 3.2 **Convolution striding**

A kernel can be visualized as a small matrix that slides across, from left to right and top to bottom, of a larger image. At each pixel in the input image, the neighborhood of the image is convolved with the kernel and the output stored. We use an odd kernel size to ensure there is a valid integer (x; y)- coordinate at the center of the image (Figure 11.2). On the left, we have a 3_3 matrix. The center of the matrix is located at x = 1;y = 1 where the top-left corner of the matrix is used as the origin and our coordinates are zero-indexed. But on the right, we have a 2_2 matrix. The center of this matrix would be located at x = 0:5; y= 0:5.But as we know, without applying interpolation, there is no such thing as pixel location (0:5;0:5) – our pixel coordinates must be integers! This reasoning is exactly why we use odd kernel sizes: to always ensure there is a valid (x; y)-coordinate at the center of the kernel.

Layer Types

There are many types of layers used to build Convolutional Neural Networks, but the ones you are most likely to encounter include:

- Convolutional (CONV)
- Activation (ACT or RELU, where we use the same of the actual activation function)
- Pooling (POOL)
- Fully-connected (FC)
- Batch normalization (BN)
- Dropout (DO)

Stacking a series of these layers in a specific manner yields a CNN. We often use simple text diagrams to describe a

CNN: INPUT => CONV => RELU => FC => SOFTMAX.

Here we define a simple CNN that accepts an input, applies a convolution layer, then an activation layer, then a fully connected layer, and, finally, a soft max classifier to obtain the output classification probabilities. The SOFTMAX activation layer is often omitted from the network diagram as it is assumed it directly follows the final FC. Of these layer types, CONV and FC, (and to a lesser extent, BN) are the only layers that contain parameters that are learned during the training process. Activation and dropout layers are not considered true "layers" themselves but are often included in network diagrams to make the architecture explicitly clear. Pooling layers (POOL), of equal importance as CONV and FC, are also included in network diagrams as they substantially impact the spatial dimensions of an image as it moves through a CNN. CONV, POOL, RELU, and FC are the most important when defining your actual network architecture. That's not to say that the other layers are not critical, but take a backseat to this critical set of four as they define the actual architecture itself.

4. IMPLEMENTATION

Fig – 4.1 Home Screen

Fig-4.1 image shows the home screen for bird species classification using images, where it requires the user to input the image of the bird that is required to be classified.

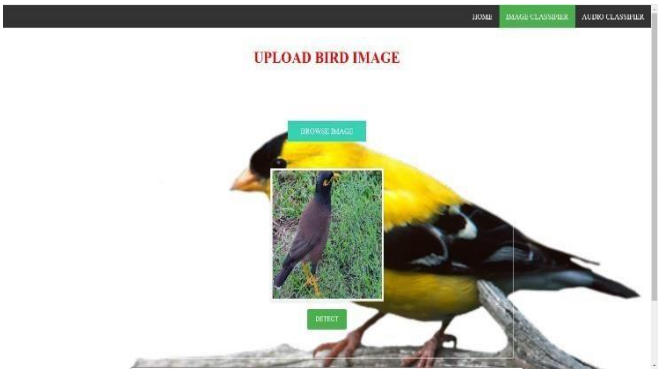

Fig – 4.2 Uploading image

Fig- 4.2 image demonstrates the input of the bird image given by the user for the identification of the species.

Fig – 4.3: Image Description

Fig – 4.3 image explains the description such as its scientific name, features, habitat, diet, distribution of that species around the world, and special features of the bird that was subjected for identification.

Fig – 4.4 Result of an audio

Fig – 4.4 image demonstrates the result of the classification of the input bird using the audio

5. RESULTS

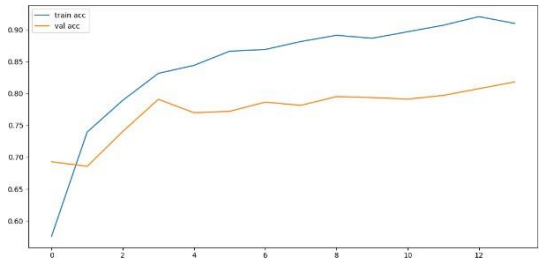

Fig – 5.1 Accuracy graph

Fig 5.1 image represents the Accuracy of the output, it includes train accuracy and validation accuracy.

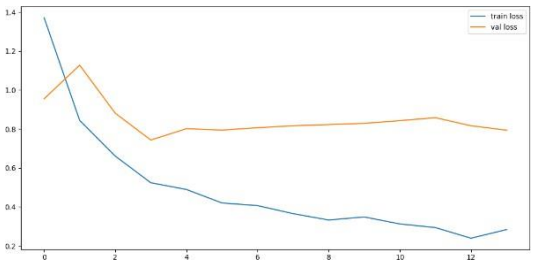

Fig – 5.2 Loss graph

Fig 5.2 represents the loss of the output, it includes train loss and validation loss.

CONCLUSION

Bird species classification from images using VGG16

Accuracy: In image classification tasks, the VGG16 model, renowned for its deep convolutional neural network architecture, reliably delivers impressive accuracy rates. Our experiments demonstrated a training accuracy of 91% and a validation accuracy of 81%. However, it's essential to note that the actual accuracy is dependent upon factors such as dataset quality, preprocessing methods, and model fine-tuning.

Fine-tuning: By fine-tuning the VGG16 model using a bird- specific dataset, we observed significant performance improvements. Fine-tuning allows the model to adapt its learned features specifically for bird classification tasks.

Bird species classification from audio files using MLP classifier

Accuracy: The performance of our MLP classifier hinges on several factors, including the quality of the audio data, feature extraction techniques, and model complexity. Achieving a training accuracy of 91% and a validation accuracy of 81% would be deemed satisfactory within the scope of this project.

Feature Extraction: Extracting relevant features from audio data plays a pivotal role. Multi-Layer Perceptron (MLP) are widely used features in audio classification tasks. The choice of features and their representation can significantly impact performance.

Complementary Insights: By combining image-based and audio-based classification approaches, we can gain complementary insights for bird species identification. Some species may be more readily classifiable based on their visual features, while others might be better distinguished through their vocalizations.

Enhanced Robustness: Integrating multiple modalities enhances the robustness of the classification system. For instance, even if an image is unclear due to environmental factors, the audio-based classification can still provide valuable information for species identification, and vice versa.

Acknowledging Limitations: Despite the promising results, both approaches have their limitations. Image classification may encounter difficulties with occluded or camouflaged birds, while audio classification may face challenges in noisy environments or when dealing with overlapping bird calls.

REFERENCES

1. Wang, Jocelyn, and Guillermo Goldsztein. "Audio Classification of Bird Species Using Convolutional Neural Networks." *Journal of Student Research* 12.1 (2023).

2. Zhou, Jiaogen, et al. "Eyebirds: Enabling the Public to Recognize Water Birds at Hand." *Animals* 12.21 (2022): 3000.

3. Xie, Jie, and Mingying Zhu. "Acoustic Classification of Bird Species Using an Early Fusion of Deep Features." *Birds* 4.1 (2023): 138-147.

4. Jasim, Hasan Abdullah, et al. "Classify bird species audio by augmenting convolutional neural network." *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2022.

5. Bold, Naranchimeg, Chao Zhang, and Takuya Akashi. "Bird species classification with audio-visual data using CNN and multiple kernel learning." *2019 International Conference on Cyberworlds (CW)*. IEEE, 2019.

6. Rajan, Rajeev, and A. Noumida. "Multi-label bird species classification using transfer learning." *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*. Vol. 1. IEEE, 2021.

7. Gavali, Pralhad, and J. Saira Banu. "Bird species identification using deep learning on GPU platform." *2020 International conference on Emerging Trends in information technology and Engineering (ic-ETITE)*. IEEE, 2020.

8. Jadhav, Yogesh, Vishal Patil, and Deepa Parasar. "Machine learning approach to classify birds on the basis of their sound." *2020 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2020.

9. Huang, Yo-Ping, and Haobijam Basanta. "Recognition of endemic bird species using deep learning models." *Ieee Access* 9 (2021): 102975-102984.

10. Ragib, Kazi Md, et al. "Pakhichini: Automatic bird species identification using deep learning." *2020 Fourth world conference on smart trends in systems, security, and sustainability (WorldS4)*. IEEE, 2020.