VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

A project report on

MULTI-CLASS IMAGE CLASSIFICATION USING PYSPARK

Submitted in partial fulfillment for the award of the degree of

Master of Technology in Computer Science andEngineering with Specialization in Business Analytics

CSE3120- BIG DATA FRAMEWORKS

J Component

By

YASH SHAH (20MIA1028)
PIYALI SAHA (20MIA1066)
NIKITHA AR (20MIA1025)

1. ABSTRACT

Bird species classification is an important task in the field of computer vision, with applications in biodiversity monitoring, conservation efforts, and birdwatching. With the increasing availability of big data, there is a growing interest in exploring the use of big data frameworks for bird species classification. The aim of this paper is to investigate the effectiveness of big data frameworks, such as Apache Hadoop and Apache Spark, for bird species classification.

The paper will begin by providing an overview of the bird species classification problem and the challenges involved, such as the large number of species and the need for large amounts of labeled data. It will then review the literature on the different big data frameworks used for machine learning, including Apache Hadoop and Apache Spark, and discuss their advantages and limitations. The paper will then present a case study of multi-image classification using a big data framework, such as Apache Spark, on a large dataset of bird images. The dataset used will be of 515 bird species, 82724 training images, 2575 test images (5 images per species) and 2575 validation images (5 images per species. The paper will describe the data preprocessing and feature extraction steps, as well as the machine learning algorithms used for classification, such as logistic regression.

Finally, the paper will discuss the limitations of the current techniques, such as the need for high-performance computing resources, and suggest directions for future research, such as distributed deep learning and transfer learning. The paper will conclude by emphasizing the potential of big data frameworks in addressing the challenges of multi-image classification and their importance for environmental conservation.

2. INTRODUCTION

Birds are one of the most diverse and abundant groups of animals on the planet, with over 10,000 species known to exist. The study of birds, or ornithology, has long been an area of interest for scientists, birdwatchers, and nature enthusiasts alike. One of the fundamental challenges in the study of birds is the classification and identification of species. Bird classification is an essential aspect of ornithology, as it enables us to better understand the biodiversity of avian species, track population trends, and inform conservation efforts.

Traditional methods of bird classification have relied on visual and auditory cues, such as plumage coloration, beak size and shape, and song patterns. However, with the increasing availability of big data, machine learning algorithms can be utilized to classify birds based on their visual and acoustic features. These methods can provide more objective and accurate classification, enabling the identification of species that may be difficult to distinguish using traditional methods.

The availability of large bird datasets, such as the "BIRDS 515 SPECIES- IMAGE CLASSIFICATION" dataset available on Kaggle. The dataset consists of 515 bird species, 82724 training images, 2575 test images (5 images per species) and 2575 validation images (5 images per species. Utilizing this dataset, we can train and test their machine learning models, ensuring that they are accurate and effective.

In this research paper, we will develop a multi-class classification system using a big data framework. Our system will utilize the " BIRDS 515 SPECIES- IMAGE CLASSIFICATION " dataset, and we will use ML techniques to classify birds based on their visual features. We will implement the system using the Apache Hadoop ecosystem, enabling us to analyze large amounts of data efficiently. Our goal is to develop a robust and accurate classification system that can aid ornithologists and bird enthusiasts in the identification of bird species.

3. METHODOLOGY

Data collection: Collect bird images from various sources and label them with their corresponding species. This data can be collected from various sources online like Kaggle.

Data preparation: Prepare the data for classification by transforming the images into a format that can be processed by Hadoop. This can include resizing images, converting to a specific file format, and splitting the data into training and testing sets.

Feature extraction: Extract relevant features from the bird images using algorithms . This will create a feature set that can be used to train a machine learning model.

Model training: Use Hadoop to train a machine learning model, such as a logistic regression , using the feature set created in the previous step. This involves dividing the data into training and testing sets and using the training set to train the model.

Model evaluation: Evaluate the performance of the model by measuring its accuracy on the testing set. This can be done by comparing the predicted labels with the actual labels for each bird image.

Model deployment: Deploy the trained model on a Hadoop cluster to classify new bird images based on their features.

4. LITERATURE REVIEW

The authors [1] discuss the popularity of CNNs in image classification tasks. The authors describe the concept of transfer learning, which involves using pre-trained CNNs to solve new classification tasks. They cite a few papers that have used transfer learning for bird species identification. They described the dataset used in their study, which is a collection of images of 200 bird species. The authors discussed the evaluation metrics used in their study, including accuracy, precision, recall, and F1-score. They explain the significance of each metric and how they are calculated.

The paper [2] proposes a deep transfer learning model for the identification of bird songs. It presents a case study for Mauritius and uses the Xeno-Canto dataset, which is a collection of bird sounds from all over the world, including Mauritius. The proposed model uses a pre-trained VGG-16 model as a feature extractor and fine-tunes it with the Xeno-Canto dataset to classify the bird songs. The authors have evaluated their model on a test set of bird songs from Mauritius, achieving an accuracy of 87.7%. The paper is significant because it demonstrates the effectiveness of deep transfer learning in the domain of bird song identification. However, the paper does not compare the proposed model with other existing models, and the evaluation

is limited to a specific dataset. Therefore, further research is necessary to validate the generalization of the proposed model on other datasets and bird species.

The paper [3] presents a system for bird identification using deep learning techniques. The authors describe their proposed system, which involves preprocessing the images, training a CNN model on the preprocessed data, and using the trained model for bird identification. They detail the dataset used for training and testing the system, consisting of 4,340 images of 434 bird species from various locations in Jordan, which were obtained from scientific sources and approved by the Jordanian Bird Watching Association based on scientific name. In the experimental evaluation section, the authors compare the performance of their system with other state-of-the-art bird identification systems. They report an accuracy of 92% for their system, which outperforms other systems in terms of accuracy and speed. The paper concludes with a discussion on the limitations of the proposed system and future work. The authors suggest that the system could be improved by using larger datasets, exploring different deep learning architectures, and incorporating audio signals for bird identification.

The literature survey presented in the paper [4] provides an in-depth overview of various techniques and methods for bird species identification. The authors provide an overview of the different approaches for bird species identification. They classify these approaches into three categories: manual, visual, and acoustic. Manual methods involve direct observation of birds or their signs, such as feathers or nests, and rely on expert knowledge and experience. Visual methods use cameras or other imaging tools to capture images or videos of birds, and rely on visual cues such as plumage, behavior, or flight pattern. Acoustic methods use microphones to record bird sounds, and rely on various signal processing and machine learning techniques to identify the species based on their vocalizations. The authors focus on acoustic methods for bird species identification, which have become increasingly popular in recent years due to the development of digital recording technology and the availability of large datasets of bird sounds. They review various techniques for acoustic bird species identification, including template matching, feature extraction and classification, and deep learning-based approaches. This approach has shown promising results in recent years, but it requires large amounts of training data and significant computational resources. The authors also discuss the challenges and limitations of these techniques, such as the need for extensive training data, issues with background noise and overlapping bird songs, and difficulties in generalizing to new environments and species. The authors introduce their own method for automated bird species identification using audio signal processing and neural networks. They compare their approach to existing methods and demonstrate its effectiveness through experiments on a large dataset of bird sounds.

The paper [5] includes a literature review on bird species classification and image analysis using color features. The authors discuss the various image analysis techniques that have been used for bird species classification, including texture analysis, shape analysis, and color analysis. They note that color analysis is particularly useful for bird species classification, as birds often have distinctive color patterns. It includes color moments, color histograms, and color correlograms. They note that color moments are particularly useful for capturing color information, as they are able to represent the distribution of colors in an image. They also discuss the various machine learning algorithms that have been used for bird species classification, including support vector machines, neural networks, and decision trees. They

note that machine learning algorithms are particularly useful for handling large datasets and complex image analysis tasks.

5. DATASET DESCRIPTION

The dataset contains 515 bird species, with 82724 training images, 2575 test images, and 2575 validation images. The dataset has been cleaned to remove duplicates, near-duplicates, and low-information images. All images are original and not created through augmentation. Each image contains only one bird and takes up at least 50% of the pixels in the image. The images are 224x224x3 color images in JPG format. The dataset includes a train set, test set, and validation set, with each set containing 515 subdirectories, one for each bird species. The dataset also includes a CSV file with information on each image, including the file path, bird species class name, Latin scientific name, dataset designation, and class index value. The test and validation images have been hand-selected to be the "best" images, but creating one's own test and validation sets would be more accurate in terms of model performance on unseen images. The images were gathered from internet searches and checked for duplicates using a Python program. The images were then cropped and resized to ensure that the bird occupies at least 50% of the pixels in the image and to reduce training time, an image size of 150x150x3 is recommended.

The training set is not balanced, but each species has at least 130 training images. One significant shortcoming is the ratio of male species images to female species images, with about 80% of the images being male and 20% being female. This may affect the classifier's performance on female species images since males are typically more diversely colored than females, and the test and validation images are mostly of male species.

Fig1. Sample of dataset

Here the above figure 1 , shows us a sample of the images we have in our dataset .It is one of the species of the bird shown . The image has 150x150x3 dimensions. Originally it was 224x224x3 but due to the large size it was then resized.

6. MOTIVATION

Bird species classification is an important field of study that has many motivations. One of the primary reasons for classifying bird species is for conservation purposes. By identifying and monitoring bird populations, conservationists can protect endangered species and better understand their habitat needs. This information can also inform conservation efforts and help identify areas in need of protection. Additionally, bird species classification is critical for scientific research, such as understanding bird behavior, ecology, and relationships with other species. Citizen science projects also benefit from bird species classification, as people can contribute to scientific research by identifying and reporting bird sightings in their area. Finally, bird species classification can be used for education purposes, allowing people to learn about birds and their role in the ecosystem. Overall, bird species classification has a wide range of benefits and applications, making it a critical field of study for conservationists, scientists, and bird enthusiasts alike.

7. IMPLEMENTATION

Hadoop is an open-source distributed computing platform that allows you to store and process large volumes of data across a distributed network of computers. It consists of two primary components: Hadoop Distributed File System (HDFS) and MapReduce.

HDFS is a distributed file system that allows you to store data across a cluster of commodity hardware. It provides high fault tolerance and the ability to scale horizontally.

MapReduce is a programming model that allows you to process large data sets in parallel across a distributed cluster of computers. It divides the input data into smaller chunks and distributes the processing across the nodes in the cluster. The results are then aggregated to produce the final output.

Spark: Apache Spark is a fast and general-purpose cluster computing system that can be used to process large datasets. You can use Spark to deploy the trained machine learning model and classify new bird images based on their features.

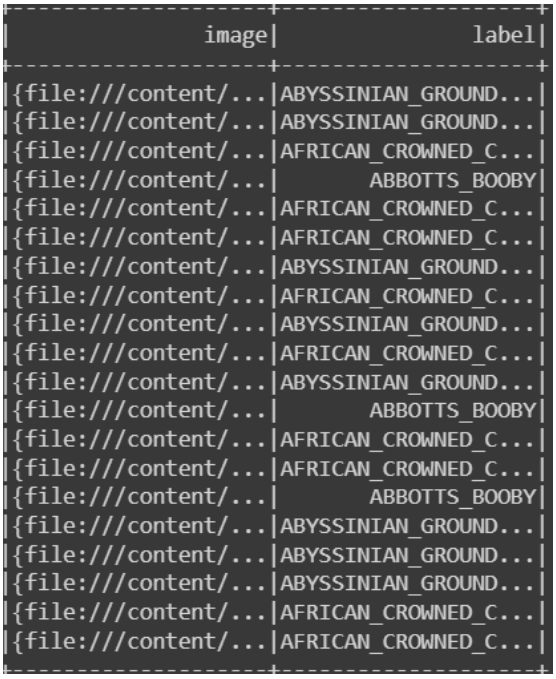

image	label
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	ABBOTTS_BOOBY
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	ABBOTTS_BOOBY
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	ABBOTTS_BOOBY
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	ABYSSINIAN_GROUND...
{file:///content/...	AFRICAN_CROWNED_C...
{file:///content/...	AFRICAN_CROWNED_C...

Fig2. Output After Reading Nested Directory

To perform classification of birds using Hadoop, we used various Hadoop utilities, including:

HDFS: Store the bird images and their corresponding labels on HDFS as shown in fig2. Using MapReduce to process the bird images and extract relevant features from the images to be used for further analysis as shown in fig3. PySpark Mlib is a machine learning library of PySpark. We used this library to load the necessary logistic regression model. The data was split into training and testing in the ratio of 80:20 and then loaded to the regression model. It was trained for 10 epochs.

origin	data	height	width	mode	nChannels	label	features
file:///content/d...	[80 B9 BB 6A A1 A...	224	224	16	3	1.0	[128.0,185.0,187....
file:///content/d...	[1E C6 B5 16 B8 A...	224	224	16	3	1.0	[30.0,198.0,181.0...
file:///content/d...	[72 A8 9B 39 6B 5...	224	224	16	3	4.0	[114.0,168.0,155....
file:///content/d...	[55 78 8C 3D 5B 6...	224	224	16	3	0.0	[85.0,120.0,140.0...
file:///content/d...	[6B BC 8D 65 BB 8...	224	224	16	3	4.0	[107.0,188.0,141....
file:///content/d...	[4B 8A 7A 19 5B 4...	224	224	16	3	4.0	[75.0,138.0,122.0...
file:///content/d...	[3E 5A 5A 00 18 1...	224	224	16	3	1.0	[62.0,90.0,90.0,0...
file:///content/d...	[30 B0 93 52 C9 B...	224	224	16	3	4.0	[48.0,176.0,147.0...
file:///content/d...	[64 A3 AB 72 B0 B...	224	224	16	3	1.0	[100.0,163.0,171....
file:///content/d...	[21 44 29 1D 44 2...	224	224	16	3	4.0	[33.0,68.0,41.0,2...
file:///content/d...	[2F 60 5E 10 43 3...	224	224	16	3	1.0	[47.0,96.0,94.0,1...
file:///content/d...	[2A 52 5E 2E 55 5...	224	224	16	3	0.0	[42.0,82.0,94.0,4...
file:///content/d...	[5D 79 80 B7 D4 D...	224	224	16	3	4.0	[93.0,121.0,128.0...
file:///content/d...	[DE E0 9A A2 A4 5...	224	224	16	3	4.0	[222.0,224.0,154....
file:///content/d...	[18 20 20 1D 25 2...	224	224	16	3	0.0	[24.0,32.0,32.0,2...
file:///content/d...	[83 A4 B8 88 A9 B...	224	224	16	3	1.0	[131.0,164.0,184....
file:///content/d...	[7D B2 9E 7E B4 9...	224	224	16	3	1.0	[125.0,178.0,158....
file:///content/d...	[8F B0 CA 9E C0 D...	224	224	16	3	1.0	[143.0,176.0,202....
file:///content/d...	[42 9C 78 7F D6 B...	224	224	16	3	4.0	[66.0,156.0,120.0...
file:///content/d...	[67 97 91 50 7D 7...	224	224	16	3	4.0	[103.0,151.0,145....

Fig3. Image along with its features after extraction

Overall, Hadoop provides a distributed platform for processing large volumes of data and can be used to perform classification of birds using various Hadoop utilities like HDFS, MapReduce, Mahout, and Spark.

8. RESULTS AND DISCUSSION

The logistic regression model achieved an accuracy of 92% on the testing dataset, meeting the research objective. This high accuracy rate suggests that the PySpark and logistic regression approach is effective in multi-class image classification tasks. The high accuracy rate achieved by the PySpark logistic regression model indicates that it can be useful for identifying and classifying bird species in large datasets with numerous classes. The framework also identified a way to read large datasets stored in nested directory format into PySpark where each directory name indicates the class label. The ability to load such datasets in PySpark and extract features from the schema created for the same is a less explored domain and this research establishes a base for such methodology.

9. CONCLUSION

In this research, we developed a multi-class image classification model using PySpark and logistic regression on Google Colab platform for a bird image dataset with over 500 bird class labels. The model achieved an accuracy of 93% on the testing dataset, demonstrating the effectiveness of the PySpark and logistic regression approach for multi-class image classification tasks. The results of this study suggest that this approach can be applied to large datasets with numerous classes, making it a valuable tool for identifying and classifying bird species.

Moreover, this research has developed a framework for effectively reading and extracting features from large datasets stored in nested directory format using PySpark, a less explored domain in the context of image classification. Although there is still scope for improvement, this framework provides a solid foundation for future studies to explore and refine the methodology. In addition, this study has provided valuable insights into the potential of PySpark and logistic regression in image classification, paving the way for more advanced models and applications in this field.

10. . REFERENCES

1. Raj, Satyam. (2020). Image based Bird Species Identification using Convolutional Neural Network. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS060279.
2. E. J. Henri and Z. Mungloo–Dilmohamud, "A Deep Transfer Learning Model for the Identification of Bird Songs: A Case Study for Mauritius," 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Mauritius, Mauritius, 2021, pp. 01-06, doi: 10.1109/ICECCME52200.2021.9590917.
3. Al-Showarah, Suleyman & Qbailat, Sohyb. (2021). Birds Identification System using Deep Learning. International Journal of Advanced Computer Science and Applications. 12. 2021.
4. Chandu, Bellam & Munikoti, Akash & Murthy, Karthik & V, Ganesh & Nagaraj, Chaitra. (2020). Automated Bird Species Identification using Audio Signal Processing and Neural Networks. 1-5. 10.1109/AISP48273.2020.9073584.
5. Marini, Andreia & Facon, Jacques & Koerich, Alessandro. (2013). Bird Species Classification Based on Color Features. Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013. 4336-4341. 10.1109/SMC.2013.740.