Bird Species Classification Using Vision Transformers

By Sakshi Borade, Atharv Chandane, Kunal Chaturvedi, Sahil Chauhan, Chavan, Prof. Kirti Mishra
Computer Engineering
K.J.Somaiya college of engineering (of Somaiya vidyavihar university
Mumbai, India

***Abstract*—Two vision transformer models, namely ViT_B-16 and ViT_L-32, were meticulously trained and deployed for the classification of 400 different species of birds. Remarkably, the ViT_B-16 model demonstrated a testing accuracy of 95.603%, while its counterpart, the ViT_L-32 model, achieved an admirable 94.272% accuracy. The models were seamlessly integrated into a user-friendly web application using Streamlit, a popular framework for building interactive data applications. This deployment allows users, irrespective of their domain expertise, to effortlessly upload images and receive predictions regarding the depicted bird species.**

Keywords—Artificial Intelligence, Deep Learning, Vision Transformers, Image Classification, Bird Species Classification, Streamlit

INTRODUCTION

This chapter presents the implementation of a bird species classification project utilizing advanced deep learning techniques, specifically Vision Transformer (ViT) models. In an era where image recognition plays a pivotal role in various domains, accurately classifying bird species from images remains a challenging task due to the diverse visual characteristics and subtle distinctions among different species. Leveraging the power of ViT models, this project aims to develop a robust classification system capable of precisely identifying bird species based on visual cues extracted from images. The implementation encompasses a comprehensive pipeline, including data preprocessing, model training, evaluation, and deployment, with a focus on optimizing model performance and usability. Through this project, the efficacy of ViT models in addressing complex image recognition tasks and their potential applications in biodiversity monitoring and conservation efforts are explored.Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

1. Background

In the ever-evolving field of ornithology and biodiversity monitoring, the quest for efficient and accurate methods to identify and quantify various bird species has driven the development of innovative technologies. Traditional qualitative descriptions often prove inadequate in providing a comprehensive understanding of the avian world. In response to this need, the Wing Watch project emerges as a groundbreaking initiative that aims to leverage the power of Vision Transformers, an advanced class of deep learning models, to simplify bird species identification through the analysis of image features. Ornithology, the scientific study of birds, has traditionally relied on manual methods for species identification, often hindered by subjectivity and limitations in accuracy. As biodiversity monitoring becomes increasingly vital, there is a pressing need to adopt technological advancements that can enhance the precision and efficiency of bird species identification. Traditional methods fall short in capturing the nuanced characteristics that define each species, leading to the emergence of a gap between qualitative descriptions and a comprehensive understanding of avian diversity.

2. Problem Statement

The problem statement of bird species image classification revolves around the challenge of accurately discerning and categorizing various avian species depicted in images. This task entails deciphering intricate visual cues, such as plumage patterns, beak shapes, wing morphology, and coloration variations, among others, to differentiate between hundreds or even thousands of distinct bird species. Given the vast diversity and nuanced features present across avian taxa, coupled with the potential for environmental factors like lighting conditions and background clutter to confound image interpretation, the task of automated bird species classification poses formidable computational and algorithmic challenges

3. Scope

The scope of the project encompasses the development of an advanced image classification system applicable to a wide range of domains and applications. The project aims to address the complex task of accurately identifying and categorizing objects depicted in images, leveraging state-of-the-art deep learning techniques. Key components within the scope include dataset acquisition and preprocessing, model selection and architecture design, training and evaluation procedures, optimization and fine-tuning strategies,

deployment and integration into user- friendly interfaces, and comprehensive documentation and knowledge transfer. By adopting a holistic approach to image classification, the project seeks to develop robust and scalable solutions capable of advancing various fields, including computer vision, remote sensing, medical imaging, environmental monitoring, and more. The project's overarching goal is to contribute to the advancement of image understanding technologies and facilitate their widespread adoption in diverse real-world applications, ultimately fostering innovation and enhancing decision-making processes across multiple domains.

4. Objectives

The objectives of the project are delineated as follows: i. Comprehensive Augmented Database: The primary objective is to construct an extensive augmented database encompassing standard image features for approximately 400 bird species. This database serves as the foundational resource for training the Vision Transformer models. By ensuring a diverse and representative set of features, the augmented database facilitates accurate species identification, enhancing the robustness and generalization capabilities of the classification system. ii. Training Vision Transformer Models: The project aims to optimize and implement an efficient classifier layer that maximizes the utilization of pretrained Vision Transformer models. By leveraging the full capabilities of the models, including learned features and hierarchical representations, the optimized classifier layer enhances the accuracy and discriminative power of species classification. This optimization process ensures that the models can effectively translate extracted features into precise species predictions, further improving the overall performance of the system.

iii. User-Friendly Web Application: The project seeks to democratize access to its advanced bird species identification system through the development of a user-friendly web application. Utilizing Streamlit, the interface will offer an intuitive platform for users to interact with the classification system effortlessly. By enabling users to upload bird images, select desired Vision Transformer models, and receive prompt and visually appealing results, the web application enhances accessibility and usability, catering to a diverse range of stakeholders, including researchers, conservationists, and citizen scientists. iv. Confidence Score and Result Display: In addition to predicting the bird species based on uploaded images, the system will provide users with a confidence score, indicating the model's certainty regarding the prediction. This feature enhances transparency and allows users to gauge the reliability of the classification results. By providing insight into the model's confidence level, users can make informed decisions based on the classification outcomes, fostering trust and confidence in the system's capabilities.

5. Methodology

To figure out how confident your model is in its prediction, we can look at the activation values from the last layer (also known as logits). These values represent how strongly the model believes in each class, but they need to be turned into probabilities to make sense to us. Here's how we can do that: 1. Start with the activation values (logits): After your model makes a prediction, you'll get a set of numbers (the activation values) for each possible class. These are raw scores, but they don't quite tell you the probability just yet. 2. Use the softmax function: To turn those raw scores into probabilities (values between 0 and 1), we apply something called the "softmax" function. It basically adjusts the scores so they sum up to 1 and shows us how likely the model thinks each class is. Here's how it works: o For each class k, we take the exponent of the activation value zk, which helps amplify larger values. o Then, we divide that by the sum of all the exponentiated scores (so the probabilities add up to 1).

3. Confidence level for the predicted class: The class with the highest probability is the one the model is most confident in. So, we just pick the class that has the highest score after applying softmax, and that's our predicted class. Finally, the confidence level is simply the probability of that predicted class: The higher this number, the more confident the model is about its prediction.

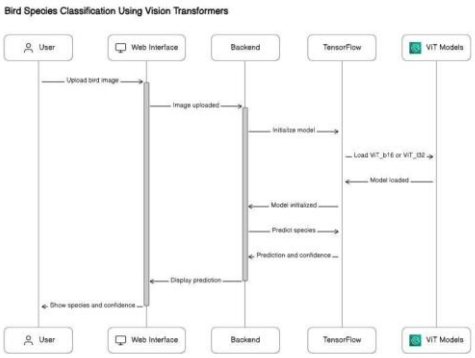
$$p_k = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}}$$

PROJECT DETAILS

6. SYSTEM OVERVIEW: THE SYSTEM OVERVIEW OF THIS PROJECT PROVIDES A HOLISTIC UNDERSTANDING OF ITS ARCHITECTURE, COMPONENTS, AND FUNCTIONALITIES. AT ITS CORE, THE PROJECT IS DESIGNED TO FACILITATE THE ACCURATE CLASSIFICATION OF BIRD SPECIES FROM IMAGES THROUGH THE INTEGRATION OF ADVANCED DEEP LEARNING TECHNIQUES AND USER-FRIENDLY INTERFACES. THE SYSTEM COMPRISES SEVERAL INTERCONNECTED MODULES, EACH CONTRIBUTING TO THE OVERALL FUNCTIONALITY AND EFFECTIVENESS OF THE BIRD SPECIES CLASSIFICATION SYSTEM. I. DATA ACQUISITION AND PREPROCESSING: THE SYSTEM BEGINS WITH THE ACQUISITION AND PREPROCESSING OF BIRD IMAGE DATASETS. THIS INVOLVES SOURCING OR CURATING A COMPREHENSIVE DATASET CONTAINING IMAGES OF VARIOUS BIRD SPECIES. THE IMAGES ARE THEN PREPROCESSED TO STANDARDIZE FORMATS, SIZES, AND ORIENTATIONS, ENSURING UNIFORMITY AND COMPATIBILITY FOR MODEL TRAINING AND INFERENCE. II. MODEL DEVELOPMENT AND

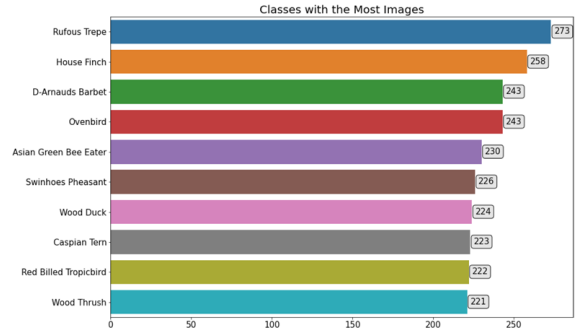
TRAINING: THE HEART OF THE SYSTEM LIES IN THE DEVELOPMENT AND TRAINING OF DEEP LEARNING MODELS FOR BIRD SPECIES CLASSIFICATION. VISION TRANSFORMER (VIT) MODELS ARE UTILIZED FOR THEIR EFFICACY IN HANDLING LARGE-SCALE IMAGE DATASETS AND CAPTURING INTRICATE VISUAL PATTERNS. THE MODELS ARE TRAINED ON THE PREPROCESSED IMAGE DATASETS, LEVERAGING TRANSFER LEARNING AND OPTIMIZATION TECHNIQUES TO ACHIEVE HIGH ACCURACY AND ROBUSTNESS.

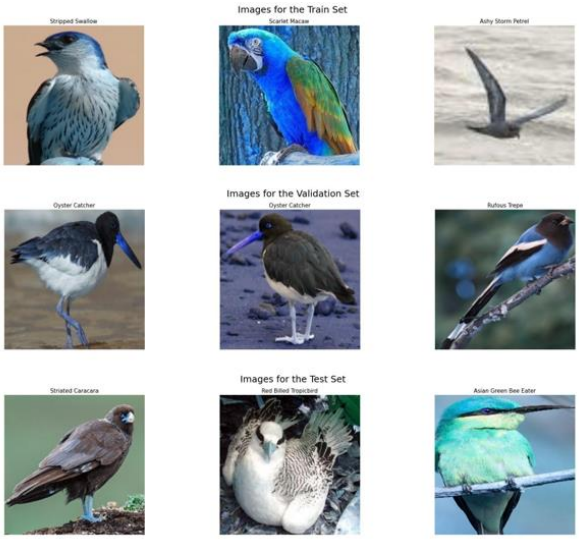
7. SOFTWARE DESIGN DOCUMENT



8. IMPLEMENTATION AND EXPERIMENTATION

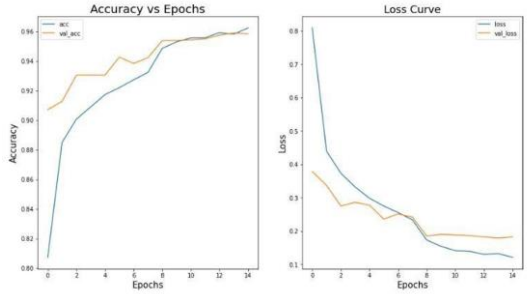
THIS CHAPTER PRESENTS THE IMPLEMENTATION AND EXPERIMENTATION PROCESS CONDUCTED TO ENHANCE THE PERFORMANCE OF VISION TRANSFORMER (VIT) MODELS FOR THE CLASSIFICATION OF BIRD SPECIES. THE UTILIZATION OF AUGMENTATIONS AND RESCALING TECHNIQUES STANDS AS A PIVOTAL ASPECT IN REFINING THE ROBUSTNESS AND GENERALIZATION CAPABILITIES OF THESE MODELS. BY SYSTEMATICALLY INTEGRATING AUGMENTATIONS AND RESCALING METHODS INTO THE TRAINING PIPELINE, WE AIMED TO BOLSTER THE MODELS' ABILITY TO ACCURATELY CLASSIFY DIVERSE BIRD SPECIES DEPICTED IN VARYING ENVIRONMENTAL CONDITIONS AND IMAGE COMPOSITIONS. THROUGH METICULOUS EXPERIMENTATION AND EVALUATION, WE SCRUTINIZED THE IMPACT OF AUGMENTATIONS AND RESCALING ON THE PERFORMANCE METRICS OF VIT_B-16 AND VIT_L-32 MODELS. SPECIFICALLY, WE MEASURED THE EFFECTS ON KEY METRICS SUCH AS RECALL SCORE, PRECISION SCORE, AND F1 SCORE, PROVIDING INSIGHTS INTO THE EFFICACY OF THESE TECHNIQUES IN IMPROVING CLASSIFICATION ACCURACY AND MODEL ROBUSTNESS. FURTHERMORE, THIS CHAPTER ELUCIDATES THE RATIONALE BEHIND THE SELECTION OF SPECIFIC AUGMENTATION TECHNIQUES AND RESCALING STRATEGIES, OUTLINING THEIR THEORETICAL UNDERPINNINGS AND PRACTICAL IMPLICATIONS. BY SYSTEMATICALLY DETAILING THE IMPLEMENTATION METHODOLOGY AND EXPERIMENTAL SETUP, WE OFFER A COMPREHENSIVE UNDERSTANDING OF THE AUGMENTATION AND RESCALING TECHNIQUES EMPLOYED TO REFINE THE PERFORMANCE OF VIT MODELS IN BIRD SPECIES CLASSIFICATION TASKS. IN ADDITION TO MODEL REFINEMENT, THIS CHAPTER ALSO DELVES INTO THE DEPLOYMENT ASPECT, WHERE THE TRAINED VIT MODELS ARE SEAMLESSLY INTEGRATED INTO A USER-FRIENDLY WEB APPLICATION USING STREAMLIT, A POPULAR FRAMEWORK FOR BUILDING INTERACTIVE DATA APPLICATIONS. THROUGH THIS DEPLOYMENT, USERS CAN UPLOAD IMAGES AND RECEIVE REAL-TIME PREDICTIONS REGARDING THE DEPICTED BIRD SPECIES, ALONG WITH CONFIDENCE LEVELS ASSOCIATED WITH EACH PREDICTION. THIS INTEGRATION NOT ONLY FACILITATES EASY ACCESS TO THE CLASSIFICATION CAPABILITIES OF VIT MODELS BUT ALSO SHOWCASES THEIR PRACTICAL APPLICABILITY IN THE DOMAIN OF ORNITHOLOGY AND BEYOND.



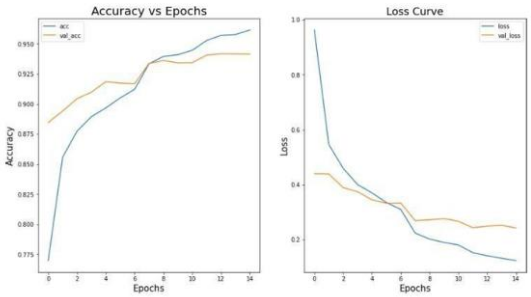


VIT_B-16: VIT_B-16, SHORT FOR VISION TRANSFORMER BASE-16, IS A SMALLER AND MORE COMPUTATIONALLY EFFICIENT VERSION OF THE VISION TRANSFORMER ARCHITECTURE. IT COMPRISES A RELATIVELY SHALLOW STACK OF TRANSFORMER ENCODER BLOCKS WITH 12 LAYERS, MAKING IT LIGHTER IN TERMS OF PARAMETERS AND COMPUTATIONS COMPARED TO LARGER VARIANTS. DESPITE ITS REDUCED COMPLEXITY, VIT_B-16 RETAINS THE ESSENTIAL SELF-ATTENTION MECHANISMS AND FEED-FORWARD NEURAL NETWORKS CHARACTERISTIC OF VISION TRANSFORMERS, ENABLING IT TO CAPTURE GLOBAL DEPENDENCIES WITHIN IMAGES AND GENERATE MEANINGFUL REPRESENTATIONS. THIS MAKES VIT_B-16 SUITABLE FOR SCENARIOS WHERE COMPUTATIONAL RESOURCES ARE LIMITED OR

WHERE A MORE LIGHTWEIGHT MODEL IS PREFERRED, WITHOUT SACRIFICING PERFORMANCE SIGNIFICANTLY.



VIT_L-32: ON THE OTHER HAND, VIT_L-32, SHORT FOR VISION TRANSFORMER LARGE-32, REPRESENTS A LARGER AND MORE EXPRESSIVE VARIANT OF THE VISION TRANSFORMER ARCHITECTURE. IT COMPRISES A DEEPER STACK OF TRANSFORMER ENCODER BLOCKS WITH 32 LAYERS, RESULTING IN A HIGHER MODEL CAPACITY AND GREATER REPRESENTATIONAL POWER. WITH ITS INCREASED DEPTH AND COMPLEXITY, VIT_L-32 CAN CAPTURE MORE INTRICATE PATTERNS AND FINER DETAILS WITHIN IMAGES, MAKING IT WELL-SUITED FOR TASKS THAT DEMAND HIGHER LEVELS OF ABSTRACTION AND SEMANTIC UNDERSTANDING. WHILE VIT_L-32 MAY REQUIRE MORE COMPUTATIONAL RESOURCES AND TRAINING TIME COMPARED TO SMALLER VARIANTS, ITS SUPERIOR PERFORMANCE AND ABILITY TO HANDLE COMPLEX DATASETS JUSTIFY ITS USAGE IN APPLICATIONS WHERE ACCURACY AND ROBUSTNESS ARE PARAMOUNT.



VIT_B-16: VIT_B-16, SHORT FOR VISION TRANSFORMER BASE-16, IS A SMALLER AND MORE COMPUTATIONALLY EFFICIENT VERSION OF THE VISION TRANSFORMER ARCHITECTURE. IT COMPRISES A RELATIVELY SHALLOW STACK OF TRANSFORMER ENCODER BLOCKS WITH 12 LAYERS, MAKING IT LIGHTER IN TERMS OF PARAMETERS AND COMPUTATIONS COMPARED TO LARGER VARIANTS. DESPITE ITS REDUCED COMPLEXITY, VIT_B-16 RETAINS THE ESSENTIAL SELF-ATTENTION MECHANISMS AND FEED-FORWARD NEURAL NETWORKS CHARACTERISTIC OF VISION TRANSFORMERS, ENABLING IT TO CAPTURE GLOBAL DEPENDENCIES WITHIN IMAGES AND GENERATE MEANINGFUL REPRESENTATIONS. THIS MAKES VIT_B-16 SUITABLE FOR SCENARIOS WHERE COMPUTATIONAL RESOURCES ARE LIMITED OR WHERE A MORE LIGHTWEIGHT MODEL IS PREFERRED, WITHOUT SACRIFICING PERFORMANCE SIGNIFICANTLY.

9. EXPERIMENTAL RESULTS

Model Name	F1-Score	Recall	Precision	Testing Accuracy
ViT_B-16	0.9560	0.9560	0.9598	95.6033%
ViT_L-32	0.9422	0.9427	0.9464	94.2718%

Model Name	Total Parameters	Trainable Parameters	Non-trainable Parameters
Neural Network with ViT_B-16 backbone	87,000,208	1,199,504	85,800,704
Neural Network with ViT_L-32 backbone	306,974,096	1,461,648	305,512,448

10. CONCLUSION AND DISCUSSION

IN THE REALM OF ORNITHOLOGY AND BIODIVERSITY MONITORING, THE PROJECT STANDS AS A TESTAMENT TO THE FUSION OF CUTTING-EDGE TECHNOLOGY AND ECOLOGICAL STEWARDSHIP. THE PROJECT ORIGINATED FROM THE RECOGNITION OF A CRITICAL GAP IN TRADITIONAL BIRD SPECIES IDENTIFICATION METHODS, PROMPTING THE NEED FOR A SOPHISTICATED SOLUTION THAT ALIGNS WITH THE ADVANCEMENTS IN DEEP LEARNING. THE PROBLEM STATEMENT IDENTIFIED THE LIMITATIONS OF QUALITATIVE DESCRIPTIONS, PAVING THE WAY FOR THE DEVELOPMENT OF A COMPREHENSIVE AND EFFICIENT BIRD SPECIES IDENTIFICATION SYSTEM. THE INTRODUCTION HIGHLIGHTED THE SIGNIFICANCE OF BRIDGING THIS GAP THROUGH THE UTILIZATION OF VISION TRANSFORMERS, ADVANCED DEEP LEARNING MODELS RENOWNED FOR THEIR PROWESS IN IMAGE FEATURE EXTRACTION. BY LEVERAGING THESE MODELS, THE PROJECT ASPIRES TO SIMPLIFY AND ENHANCE BIRD SPECIES IDENTIFICATION, CONTRIBUTING TO THE EVOLVING LANDSCAPE OF ORNITHOLOGICAL RESEARCH.

REFERENCES

[1] Noumida, A., et al. "Stacked Res2Net-CBAM with Grouped Channel Attention for Multi-Label Bird Species Classification." 2023 31st European Signal Processing Conference (EUSIPCO). IEEE, 2023.

[2] A. Marini, J. Facon and A. L. Koerich, "Bird Species Classification Based on Color Features," IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, pp. 4336-4341, doi: 10.1109/SMC.2013.740. keywords: {Birds;Image color analysis;Image segmentation;Feature extraction;Vectors;Visualization;Histograms;pattern recognition;color features;color image segmentation;machine learning;bird species classification}

[3] Asmita Manna, Nilam Upasani, Shubham Jadhav, Ruturaj Mane, Rutuja Chaudhari and Vishal Chatre, "Bird Image Classification using Convolutional Neural Network Transfer Learning Architectures" International Journal of Advanced Computer Science and Applications(IJACSA), 14(3), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140397

[4] Shriharsha, Tushara, Vijeth, Suraj, Hemavathi, "Bird Species Classification Using Deep Learning Approach", International Research Journal of Engineering and Technology (IRJET), Volume 07, Issue 04, April 2020.

[5] Samparthi V S Kumar, Hari Kishan Kondaveerti, "A Comparative Study on Deep Learning Techniques for Bird Species Recognition", 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), pp.1-6, 2023.

[6] https://www.kaggle.com/datasets/gpiosenka/100-bird-species

[7] https://paperswithcode.com/method/vision-transformer

[8] https://github.com/faustomorales/vit-keras

[9] https://github.com/google-research/vision_transformer

[10] https://streamlit.io/

We express our deepest gratitude to Prof. Kirit Mishra for her exceptional mentorship, invaluable guidance, and unwavering support throughout the implementation of this engineering project. Her profound expertise, patience, and dedication have been instrumental in shaping our perspectives and methodologies.