Identifying Bird Species through Machine Learning/Deep Learning

Aditi Roy
MT23010

Tarang Viroja
MT22081

1. Problem statement and motivation

Bird species identification presents a fascinating challenge due to the immense diversity in avian life and the brief differences in their appearances. Traditional methods of classification often struggle to cope with this complexity. But now that we have advanced techniques in Machine Learning (ML) and Deep Learning (DL), we can completely change the way we approach this task.

By accurately classifying bird species, researchers can better understand ecosystem dynamics, monitor biodiversity, and even aid in the conservation efforts of endangered species.

However, several challenges hinder the progress in this domain. The vast variability in bird species, coupled with complex backgrounds in natural habitats, poses obstacles for conventional classification methods.To overcome these challenges, we propose a comprehensive approach that integrates the strengths of both Machine Learning and Deep Learning techniques. By leveraging the power of these methodologies, we aim to develop a robust classification system capable of accurately identifying bird species based on their unique characteristics and categories. By harnessing the potential of ML and DL, we aspire to contribute to the advancement of avian research and conservation efforts, ultimately fostering a deeper understanding and appreciation of the rich diversity of bird life on our planet.

2. Literature review

Through the review of existing research papers in the field, it has been observed that various ML models have been proposed for bird species classification. However, to enhance accuracy and overcome the limitations of traditional methods, we propose to incorporate convolutional neural networks (CNNs) and vision transformers into our approach.

2.1. Leveraging Traditional Machine Learning Techniques:

We have begun by exploring traditional ML techniques that have been employed in previous studies for bird species classification.By extracting features and training ML classifiers, we aim to establish a baseline performance for bird species recognition.

2.2. Integration of Deep Learning:

CNNs have demonstrated remarkable success in various computer vision tasks, including image classification, due to their ability to automatically learn hierarchical features. We plan to leverage pre-trained CNN architectures (e.g., VGG, ResNet, Efficient) as feature extractors or fine-tune them on our bird species dataset to capture intricate patterns and representations specific to bird images. Along with all this, we are also integrating Vision Transformer.

3. Dataset

The dataset that is going to be used for this project is ”CUB-200-2011”.
Number of classes or bird species: 200 Total number of images: 11788 11788 images 5888 in test and train contain 5900.

link:https://www.vision.caltech.edu/datasets/cub_200_2011/

3.1. Data Preprocessing

We first pre-processed the data to make the height and width of all images equal. All the images are resized to 224x224 pixels. We normalized the data across all three dimensions by the meanof 0.485, 0.456, and 0.406, respectively. The variance was scaled by a factor of 0.229, 0.224, and 0.225 across three dimensions. We augmented the dataset by taking random crops, performing random horizontal flips, and random rotation of 20° C of the training images

4. Proposed architecture

We implemented the model architecture in PyTorch.The loss function used is the Categorical Cross Entropy Loss. We trained our architecture using 4700 images and validated it against 1200 images. We used Stochastic Gradient Descent (SGD) as our optimizer with a momentum of 0.9 for 30 epochs. Empirically, we found that SGD performs better with images compared to Adam optimizer. The initial learning rate was set to 0.001. Except for the baseline models, all the other models use a StepLR scheduler that decays the learning rate. All the models were trained on Nvidia GeForce GTX 1650Ti GPU

4.1. ML Techniques

While analyzing the dataset, we came to the conclusion that the dataset is very vast and of high dimensional space. Generally a simple image classification process can be done using two methods SVM(Support Vector Machine) and QDA(Quadratic Discriminant Analysis). As we are working on such high dimensional data, it will be very difficult to load such data for SVM and QDA due to which the processing will also take time. To ease our work, we tried to reduce the dimension of our dataset using a dimension reduction technique called PCA(Principal Component Analysis). For that, we took the first 1000 components into consideration and then calculated the accuracy for the SVM and QDA model which came out to be terrible. For PCA, it came out to be about 3

4.2. CNN Architecture (Baseline)

Convolutional neural networks, or CNNs, are a superior option for our project because of their automatic feature extraction capabilities, which allow us to retain the accuracy and efficiency of the image classification process. As CNN goes through hundreds of hidden layers on the images, it learns the features. Thus, we are using CNN based architecture like VGG, ResNet18 and AlexNet.

4.3. EfficientNet

With less parameters than conventional CNNs, the EfficientNet family of convolutional neural network (CNN) architectures is intended to provide higher performance. Compound scaling is the main technique of EfficientNet, which involves methodically increasing the depth, width, and resolution of CNNs. While maintaining the other dimensions unchanged, traditional techniques frequently scale one of these dimensions. To find the best balance between model size and accuracy

4.4. Vision Transformer

In contrast to conventional convolutional neural networks (CNNs), the vision transformer is a revolutionary method for image classification and other computer vision problems. In order to produce a series of feature vectors, the vision transformer's basic method involves splitting the input image into smaller patches, which are subsequently flattened and linearly embedded. After that, these feature vectors are input into a transformer architecture, which makes use of self-attention processes to identify relationships and global interdependence among various patches in the image. The vision transformer directly analyzes the image as a series of patches, which enables it to capture long-range relationships more successfully than CNNs, which rely on hierarchical feature extraction through convolutional layers. Utilizing self-attention mechanisms, the vision transformer is able to effectively simulate intricate relationships between patches and adapt

Model Name	f1 Score	Accuracy	Precision Score	Recall Score
VGG	0.6066	0.6149	0.6514	0.6177
Resnet18	0.3020	0.3310	0.3480	0.3345
AlexNet	0.4174	0.4249	0.4487	0.4284
EfficientNet	0.7685	0.7668	0.7765	0.7697
Vision Transformer(grey)	0.6377	0.6418	0.6651	0.6440
Vision Transformer	0.8034	0.8061	0.8175	0.8068

4.5. Enlighten GAN

EnlightenGAN represents a significant advancement in low-light image enhancement through its innovative use of generative adversarial networks (GANs). Unlike traditional approaches reliant on paired image data, EnlightenGAN establishes an unpaired mapping between low and normal light image spaces, reducing the need for synthetic or limited real paired data. This departure enables more flexible and efficient training, marked by shorter training times. It incorporates novel techniques such as dual-discriminators for global and local enhancement balance and illumination information for attentional guidance. These innovations facilitate unsupervised learning and enhance real-world generalization, surpassing previous methods in visual quality and no-referenced image quality assessment.

5. Visualizations

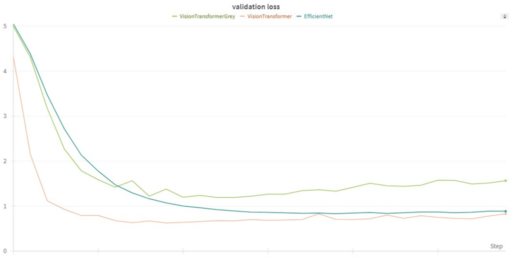

Figure 1. 'Validation loss'

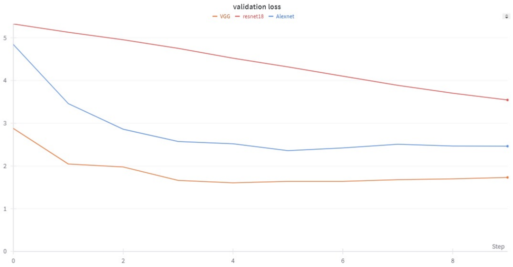

Figure 2. 'Validation loss of baseline'

6. Results

7. Analysis of Results

The table presents a comparison of model performance metrics, including f1 score, accuracy, precision score, and

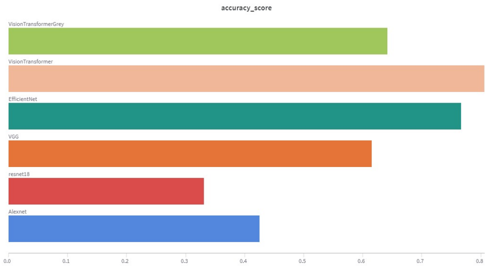

Figure 3. 'Comparison of different accuracies '

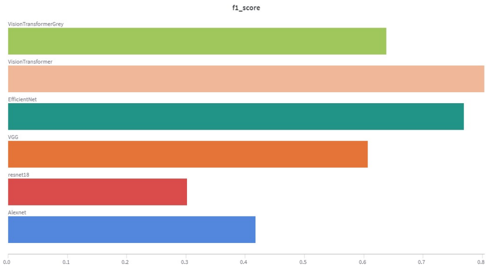

Figure 4. 'Comparison of different f1 scores '

recall score, for various deep learning models. Among the models evaluated, EfficientNet demonstrates the highest overall performance, with an f1 score and accuracy of approximately 0.77. It achieves a balanced trade-off between precision and recall, suggesting effective identification of positive cases while minimizing false positives. Following closely is the Vision Transformer model, exhibiting comparable performance to EfficientNet with an f1 score and accuracy of around 0.80. Conversely, Resnet18 performs notably poorer, displaying lower f1 score and accuracy, as well as weaker precision and recall scores. Models such as VGG and AlexNet fall in between, displaying moderate performance across the metrics. These results underscore the significance of choosing appropriate architectures for specific tasks, with EfficientNet and Vision Transformer emerging as promising candidates for tasks requiring robust performance.

8. Individual contributions of each group partner

- **Research and literature survey:** Tarang and Aditi
- **SVM, PCA and Data preprocessing:** Aditi
- **DL techniques:** Tarang
- **Data Augmentation and EfficientNet:** Aditi
- **Vision Transformer:** Tarang
- **Report:** Aditi and Tarang

9. Github Repository

Please find the source code for this project on my GitHub repository: SML Project

References

[1] Wah, C. and Branson, S. and Welinder, P. and Perona, P. and Belongie, S *WahCUB2002011*, 2011 [Online]. Available: https://authors.library.caltech.edu/records/cvm3y-5hh21

[2] Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, Vijayan K. Asari, *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1803.01164

[3] tSasha Targ, Diogo Almeida, Kevin Lyman *Resnet in Resnet: Generalizing Residual Architectures*, 2016, [Online]. Available: https://doi.org/10.48550/arXiv.1603.08029

[4] Haque, Md Foysal and Lim, Hye-Youn and Kang, Dae-Seong *Object Detection Based on VGG with ResNet Network*, 2019 [Online]. Available: https://doi.org/10.23919/ELINFOCOM.2019.8706476

[5] Tan, Mingxing and Le, Quoc *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, 2019 [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[6] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir *Transformers in Vision: A Survey*, 2022 [Online]. Available: https://doi.org/10.48550/arXiv.2101.01169

[7] Yifan Jiang and Xinyu Gong and Ding Liu and Yu Cheng and Chen Fang and Xiaohui Shen and Jianchao Yang and Pan Zhou and Zhangyang Wang*EnlightenGAN: Deep Light Enhancement without Paired Supervision*, 2021 [Online]. Available: https://doi.org/10.48550/arXiv.1906.06972