CPE 695 Final Project Proposal

Raja Agireddy Joshua Grou

Problem Statement:

For our final project we are working with the 450 bird species data set (<https://www.kaggle.com/datasets/gpiosenka/100-bird-species>). The goal will be to train machine learning algorithms such that we can correctly classify the birds into their species. As a stretch goal, we would like to attempt to predict the sex of the bird as well (https://towardsdatascience.com/bird-species-classification-with-machine-learning-914cbc0590b)

Description of the data set

The data set is composed of a collection of images of 450 species of bird. The data set has already been segmented into a training, test, and validation set, with 70,626, 2250, and 2250 images respectively. The images are stored in a folder structure that follows structure main folder -> test/train/validation -> species -> ####.jpg. This data set is imbalanced and not all of the species of bird have equal representation of in the dataset. The compiler of the data set also noted there was a lack of female representation in the dataset with about 80% of the images being male.

Tentative implementation plan, timeline, and milestones

Data Exploration -> Algorithm Selection -> Data Preparation -> Model Tuning -> Model Evaluation

In the image recognition field we have not seen extensive types of data exploration as there aren’t traditional features, we will continue to research different methods of gaining some insight into the data. From initial exploration of the image recognition field is seems that most people are using neural network type algorithms for this problem tentatively we will explore two implementations of these algorithms through the two most popular packages (pytorch and tensorflow). Some additional research will need to be placed into what if any pre-processing of the data will need to be done in addition to formatting as the package doing the learning requires. Before we can begin to tune the model we will need to choose and appropriate evaluation metric, as this is a classification task F1 score is a promising start. Further research will need to be done to understand all of the hyper-parameters involved with neural network training.

Milestone

Algorithm is defined for each member -> Code is implemented in a “ready to train state” -> Selection of hyper-parameters and evaluation metric -> Model Training and Optimization -> Prepared Report

Timeline

Data Analysis and Algorithm Selection: 11/5

Code Implementation: 11/19

Hyper-Parameter Selection and Evaluation Metric: 11/16

Model Training and Optimization: 12/03

Report Prepared: 12/10

Team members and task allocation

Raja Agireddy: Tensorflow (data prep, model coding, model tuning and optimization), lead for data exploration.

Joshua Grou: Pytorch (data prep, model coding, model tuning and optimization), lead for report preparation.