CSCI 578 Machine Learning
Competition 2022: Butterfly Classification

Instructor: Murat Dundar

Due date: 4/1/2022

Please type in your assignment and submit it online through Canvas. No hard copies will be accepted.

Task 1: Closed-set Butterfly Classification using ResNet-101 Features

Dataset: In this assignment you are given image embeddings of butterflies from over 1,000 different species. The image data set is obtained from six different museum collections. The data is split into three as train, validation, and test. Train and validation samples have 2048-dimensional vector embeddings and their matching species labels whereas testing samples have only vector embeddings. All image embeddings are obtained by a pretrained [1] ResNet-101 model. The training set contains samples from species not present in the testing set. The testing set also contains samples from species not represented in the training set. See Tables 1, 2, and 3 for information about train, validation, and test sets. See Figure 1 for sample images. Note that train/validation and test images do not come from the same set of museum collections. Train/validation images are obtained from two collections, which are randomly split into train and validation sets. Test images are obtained from four other collections.

Goal: The goal of this first task is to predict species labels for testing samples.

Approach: You are allowed to use any classifier algorithms covered in the class. SVM, Logistic Regression, and Student-t classifier are some possible alternatives. You are allowed to use any normalization and dimensionality reduction techniques.

Variable name	Size	Description	Type	
classid	1x7849	Species names as class labels	String	
features	7849x2048	ResNet-101 Features	Float	
imid	7849x1	Image IDs	Integer	
sampleid	7849x1	Specimen IDs	Integer	

Table 1: Training Dataset

[1] ResNet-101 pretrained with 1M images from ImageNet

Variable name	Size	Description	Type	
classid	1x1379	Species names as class labels	String	
features	1379x2048	ResNet-101 Features	Float	
imid	1379x1	Image IDs	Integer	
sampleid	1379x1	Specimen IDs	Integer	

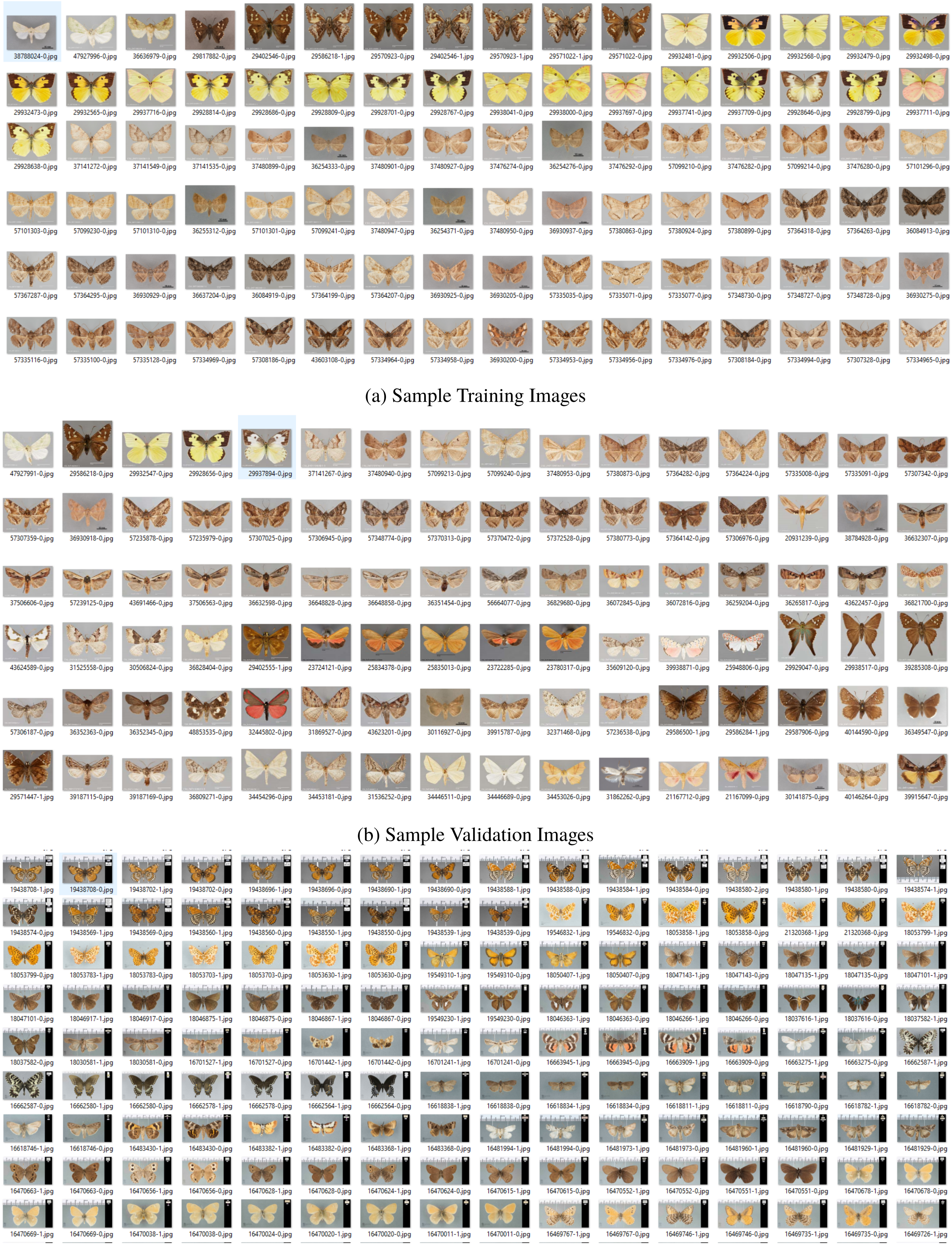
Table 2: Validation Dataset

Variable name	Size	Description	Type	
features	2454x2048	ResNet-101 Features	Float	
imid	2454x1	Image IDs	Integer	
sampleid	2454x1	Specimen IDs	Integer	

Table 3: Testing Dataset

Submission: You are going to submit a spreadsheet with predicted labels. The order of predictions should match the order of the samples in the $xtest$ matrix. You can include up to three columns in this spreadsheet, one column for each of the different approaches you have tried. You will also send a one-page PDF document explaining each approach you used in detail. This document will not be graded but if not submitted may disqualify your submission. Although you are not required to submit your code your submissions should be reproducible and your code may be requested if needed.

Evaluation: Your submission will be evaluated by mean class accuracy. For each class in the testing data true positives (TP) and false negatives (FN) will be obtained and class-level accuracy will be computed as $\frac{TP}{TP+FN}$. Mean class accuracy will be used for final evaluation and ranking. Top performer will receive the top score of 75 points. The rest of the submissions will be scored based on how well they do relative to the top performer.



(a) Sample Training Images

(b) Sample Validation Images

(c) Sample Test Images

Figure 1: Sample Images.