Midterm Checkpoint Draft

General Checklist

1. Intro/Background (From Proposal)
2. Problem Definition (From Proposal)
3. Methods
4. Results and Discussion

Methods

Preprocessing

- Eliminating Bad Features
    - Data Pruning/Cleaning
        - Getting rid of folders with small number of images
- Image Compression
    - Standardization

Steps for Standard Scaler

1. Resize and make array of all images [img1, img2,...] for all folder
2. Make array of all labels [label1, label2, …] (essentially img - label pair)
3. Use train_test_split() on array of all images to create training and testing set.
4. Use train_test_split() on training set to create training set and validation set
5. Fit standard scaler to training set
6. Transform training, validation, and testing sets using this standard scaler

    EX (step 3+):

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.15)
X_train, X_val, y_train, y_val = train_test_split(X_train, y_train, test_size = 0.2)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_val = scaler.transform(X_val)
X_test = scaler.transform(X_test)
```

Summary Text

Before our preprocessing step, we decided to first clean the dataset deleting non-uniform resolution images, and then deleted folders (classes) that had less than 100 images, as we believed it would be hard to classify images of those classes because of the small amount of data given. Lastly, due to the large dimensionality of the dataset, we decided to resize every remaining image into ⅓ of its original size.

Next, for the actual preprocessing step, we implemented standardization across the entirety of the remaining data. Using sklearn's StandardScaler(), train_test_split(), and fit_transform() methods, we first split the data into 85% training and 15% test. We then split that 20% of that training data into testing and the rest into training. Lastly, we fit the transformation onto each set, and got our standardized images for training, validation, and testing.

Machine Learning Algorithm Implemented

- Supervised Learning: Convolutional Neural Network

Results and Discussion

Results/Data Visualization

- Image to Classification Results
    - CONFUSION MATRIX
    - F1 Score
- MathPlotLib Graph Plots

Analysis of Convolutional Neural Network

Next Steps

- We plan on testing out our next model implementation, and also utilize a new preprocessing method as well.