Name: Shrey Patel
ID: 012430652

A Report on Assignment2: Image Classification

Submitted By: Shrey Patel (012430652)

Problem Statement: Develop predictive models that can determine, given an image, which one of 14 classes it is.

F1 score (as seen in CLP system, 50% data) : 0.7402
Rank: 23

Approach taken to solve the classification problem:

1. **Pre-processing**
   - Given inputs are the raw jpeg images, so first we need the convert then into vectors with uniform dimensions.
   - First, I scale the images to 64X64 dimensions (and 3 channels for RGB). For this, I use Linear Interpolation on the raw images.
   - **Feature Extraction: Histogram of Gradients:**
     - Histogram of Gradients (HOG) is a very useful image feature extraction technique, which can be used for classification purposes.
     - I applied Hog with 8 orientations, 8X8 pixels per cell, and 1X1 cells per block. These values were obtained after iterating over several different combinations, and visualizing the output image vector.
   - 2 Approaches were experimented with hog:
     a. Apply hog before re-scaling
     b. Apply hog after re-scaling

     Among this 2 approaches, option b. seemed to be more promising, as option a. resulted in some extended dashes in image visualization.
   - The resulting hog feature vector obtained had 512 features.
   - **Dimensionality Reduction:**
     - Several DR methods were applied to check the performance of end results.
     - Used singular value decomposition(ncomponents=200) initially to test the performance, but the results were no so encouraging.
     - Used PCA by first fitting the PCA model to get explained variance ratios. The values started from the order of 10^{-2} to went low till the order of 10^{-22}.
     - But a significant breakpoint was observed after first 460 values, where the order of value suddenly decreased from 10^{-6} to 10^{-16}. So, first 460 values were a good choice the experiment with.

- On applying kNN model with the PCA induced 460 features, the performance was significantly dropped from .70 F1 score to .36 F1 score (on traffic-small dataset), so I did not used PCA in this program, as the results turned no better. However, significant performance on the run time of classifier was seen, as the dimensions were reduced.

2. Classifier Algorithm

1. kNN Classifier

- Initially, used kNN with k=5 on small-traffic dataset, and got .72 F1 score, which seemed to be fair enough.
- However, runtime was a major concern, when this model was applied to large dataset of 100000 rows.
- The program kept on running entire night, which clearly indicated that kNN is not a good choice.

2. SVM

- Support Vector Machines is also a very good algorithm in binary classification.
- For this multiclass classification, I used SVM with one-on-one classification strategy for dealing with this multi class classification problem.
- SVM gave good performance of 0.71 F1-score, but took almost 1.2 hr. to learn from large dataset.

3. Random Forest

- When applying random forest on large data set, we get the F1 Score of 0.7402 on 50% of the data.
- I used 50 estimators (decision trees) with max depth of 4 levels, and Gini index was selected as the split criterion.
- Interestingly, it took only about 1 minute in training phase for the large data set, and prediction on the large Test dataset took only 0.21 minute.

4. Improving overall performance of system.

- Since the data set contained 10 million samples, applying the pre-processing step and extracting Hog features was also time consuming.
- For this, my program will save the extracted features to a text file for the first time, and read from that file only from then onwards. With this approach, the overall runtime is improved by **7 minutes.**