Laboratory 7 - PHOW Classification

Ricardo MENDOZA-LEON
Universidad de los Andes
Bogotá Colombia
ra.mendoza35@uniandes.edu.co

1. Dataset

In this lab, two datasets were used. The first one is Caltech101, which contains 102 classes totaling 3000 color images, with uneven number in each class. This are images of everyday scenes, but taken by photographers. Mayor variability is reflected in inter-class variability, being most of the images in the same class very similar, with orientation normalized using cropping and rotations.

The second dataset is a subset of imagenet-tiny dataset, which contains 200 classes totaling 20.000 color images, from Flickr. Contrary to Caltech101, this dataset has an even number of images in each class (100 in each class). Variability in this images is high; noteworthy is the intra-class variability (much more than Caltech101). Also the various scales and locations of classes, the multiplicity of instances that can appear in an image and the background complexity, make this dataset very complex and challenging.

2. Methods

The base recognition pipeline PHOW_CALTECH101 implemented by Andrea Vedaldi, is very similar to the one described in the lab6. The mayor stages are:

A) Dictionary creation: After image setup, this method, instead of using a bank of filters, the base descriptor is the Pyramid Histogram of Oriented Gradients (PHOG), that calculates histograms of gradient orientations for a number of regions inside a window centered at regular steps in the image. Given that is a multi-scale model (pyramid), this process divides the window along the x and y axis (spatial divisions $sDivX$, $sDivY$) at different heights or $levels$; the number of orientation bis are usually 8. With this procedure, a feature vector is extracted at each window location. Then, features are quantized to a number of given words k. In Andrea's pipeline, quantization is obtained by using k-means, then creates a space division tree (KD-Tree) as a lookup structure which offers a very fast way to label new vectors (by traversing the tree ideally in log(k) steps).

B) Classifier training and prediction: A set of binary classifiers is trained to recognize examples of each class. The classifiers used are SVMs with a Stochastic Dual Coordinate Ascent SDCA kernel. Later, in the prediction step given a image representation, the class corresponding to the maximum confidence value for all classifiers is used to label the image. For tests in imagenet-tiny we used 90% of images for training and 10% for test.

3. Results and discussion

A first test was run using the caltech101 dataset with an equal number of images for training and test (15/15), using the default settings of Andrea's code (k =600 and two level PHOG with $sDivX = sDivY = [2, 4]$) for 102 classes. The score matrix for train and test images, and the confusion matrix for the classification task are shown in figures 1 and 2, were it can be observed a global accuracy of 68.10%. However, it can be noticed in the confusion matrix some classes whose accuracy is very low (dark blue values at the diagonal), but its corresponding value in the train scores matrix is high, meaning over-fitting occurred for that class classifier. Noteworthy are the first five classes in the dataset that have almost 100% accuracy.

Different configurations of dictionary and spatial partitioning were tested on imagenet-tiny dataset. Figure 3 and 4, shows the results for imagenet-tiny dataset with parameters k =600 and two level PHOG with $sDivX = sDivY = [2, 4]$ in 200 classes. The first thing to notice is the low accuracy obtained for this dataset, which is caused by the much higher variability, but also the higher granularity of the classes. For instance, the chihuahua class had 0% accuracy as shown in figure 5. Furthermore, confusion and score matrices (figure 6) reveal misclassification with the class English_springer (figure 8), which is and categorically close to Chihuahua (figure 7).

An additional set of parameters: k =1200 and $sDivX = sDivY = [2, 4]$ in 200 classes, and k =600 with $sDivX = sDivY = [4, 8]$ in 50 classes, was tested on the imagenet-tiny (figures 9 and 10, and 11 and 12), which improved the accuracy to 28.65% and 37.40% respectively.

An analysis of the effects of increasing the dictionary

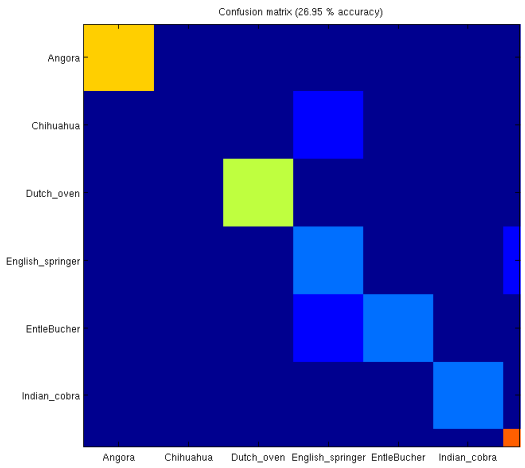

Figure 5.

size and a finer spatial partitioning on the Chihuahua class, showed no improvement for this class as seen the confusion and score matrices (figures **??** and 14, and **??** and 16). However, we note the inclusion of the EntleBucher (another breed of dog) class **??**. In conclusion the PHOG descriptor was incapable of representing the distinctive features for class Chihuahua dog, nonetheless the fact that misclassification are related with the same categorization (dogs) indicates that the model is extracting some amount of the structural features of dogs. Similar results can also be seen with other classes in this dataset that are hindering further improvements in accuracy.

As possible improvements it is proposed: a) the inclusion of additional texture information (e.g. textons), b) color features and c) spatial and region information, using the result of a segmentation algorithm such as gPB-OWT-UCM.

Figure 7.

Figure 8.

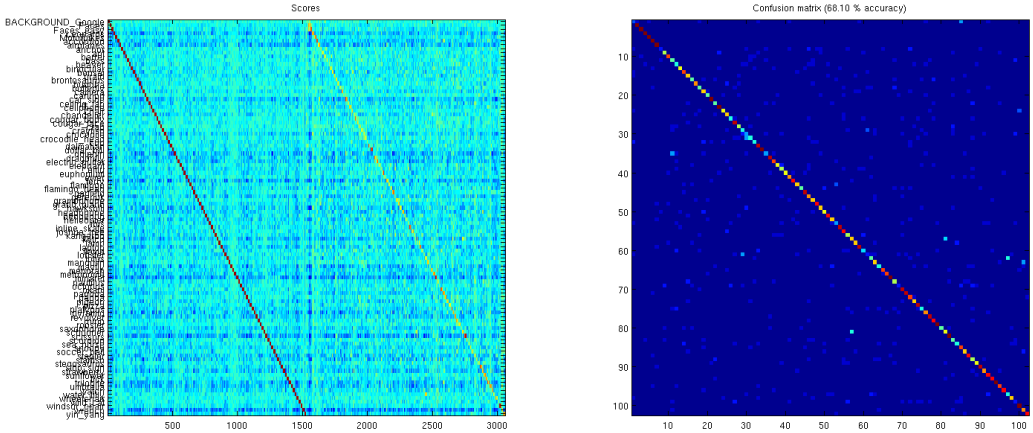

Figure 1.

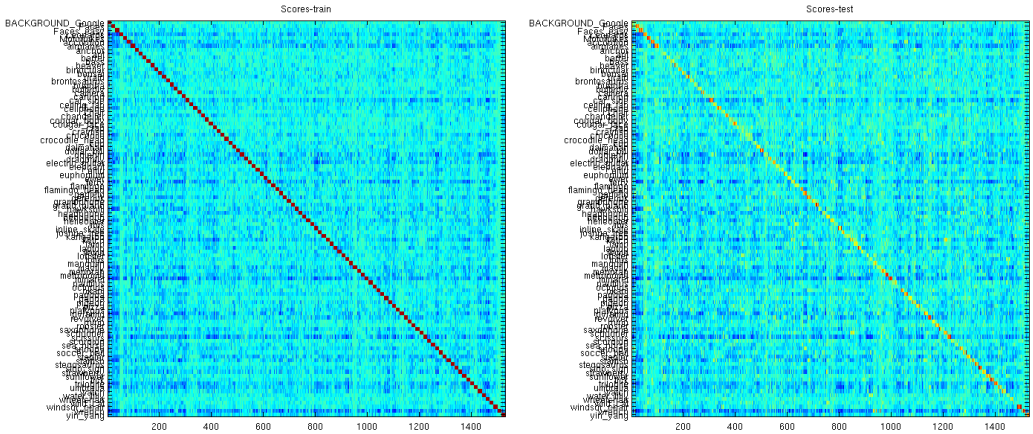

Figure 2.

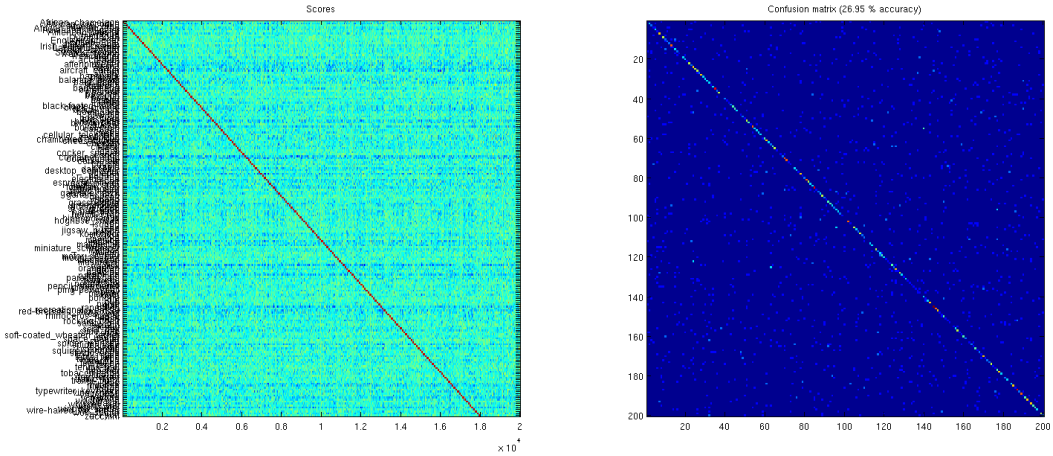

Figure 3.

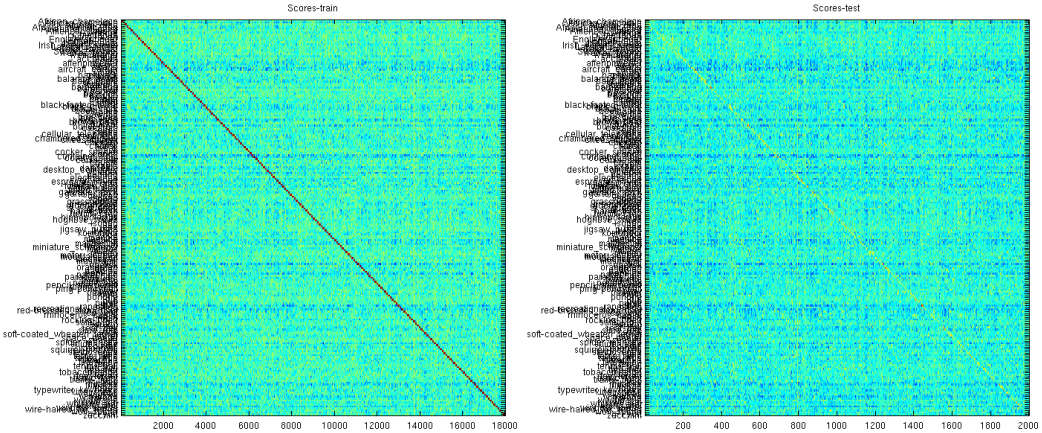

Figure 4.

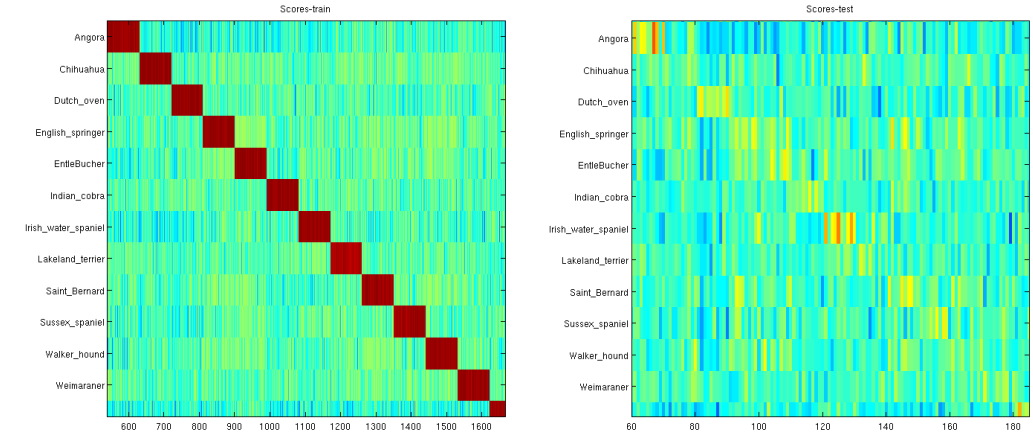

Figure 6.

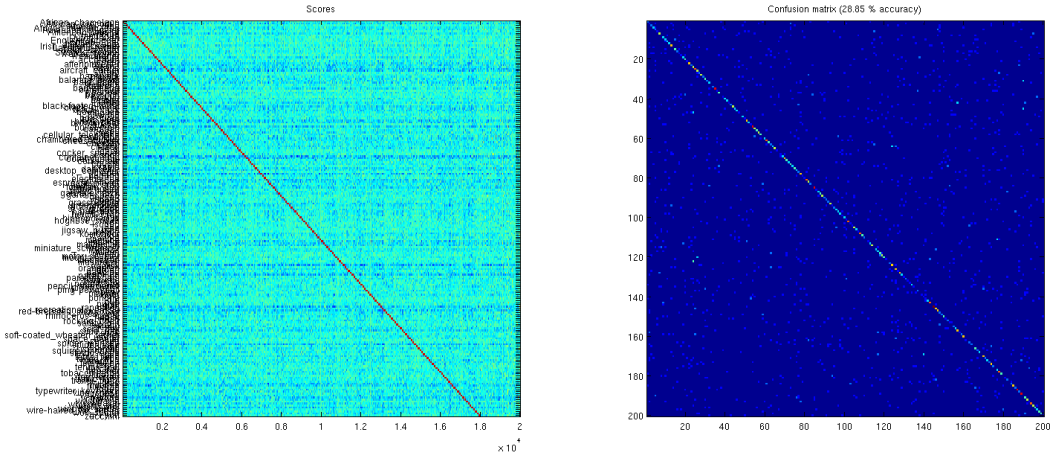

Figure 9.

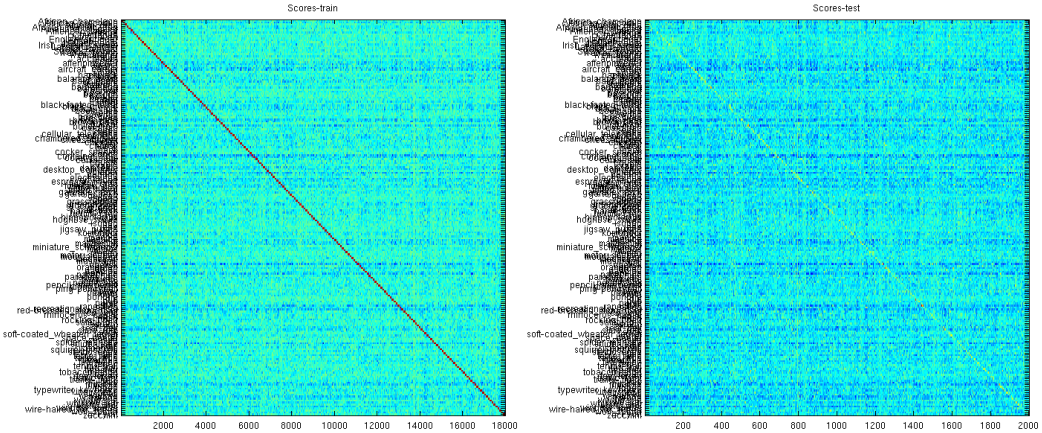

Figure 10.

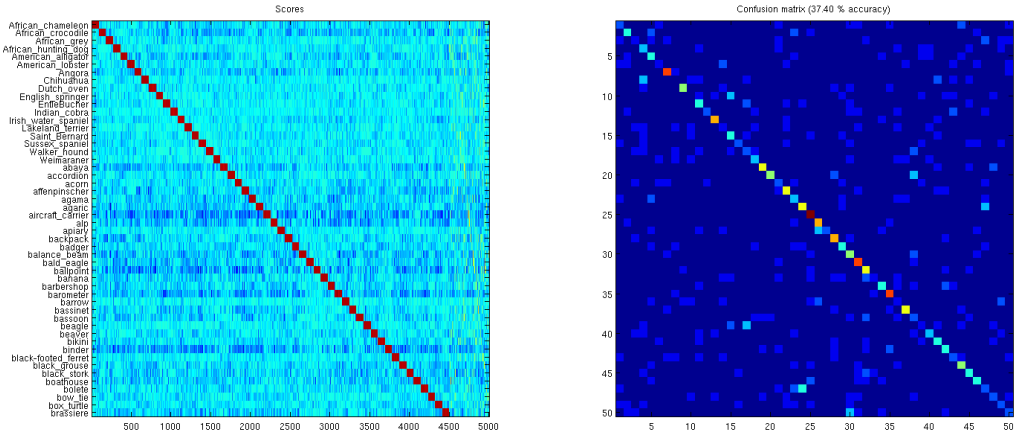

Figure 11.

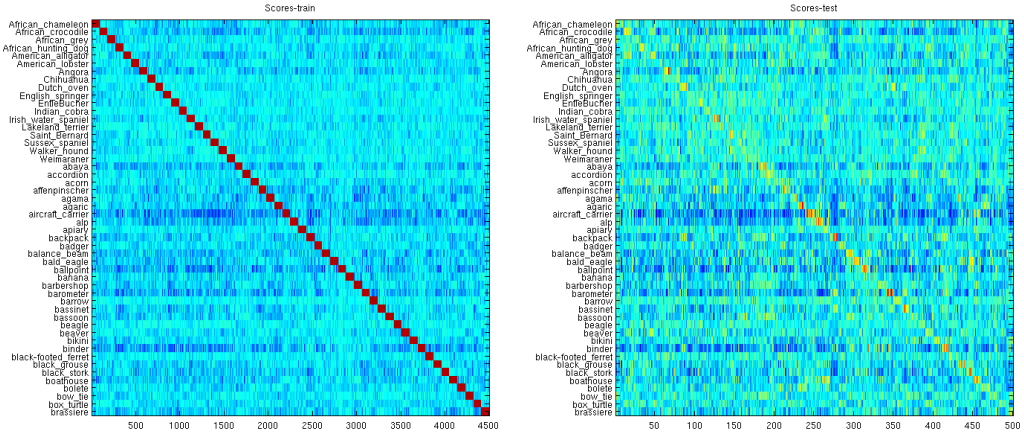

Figure 12.

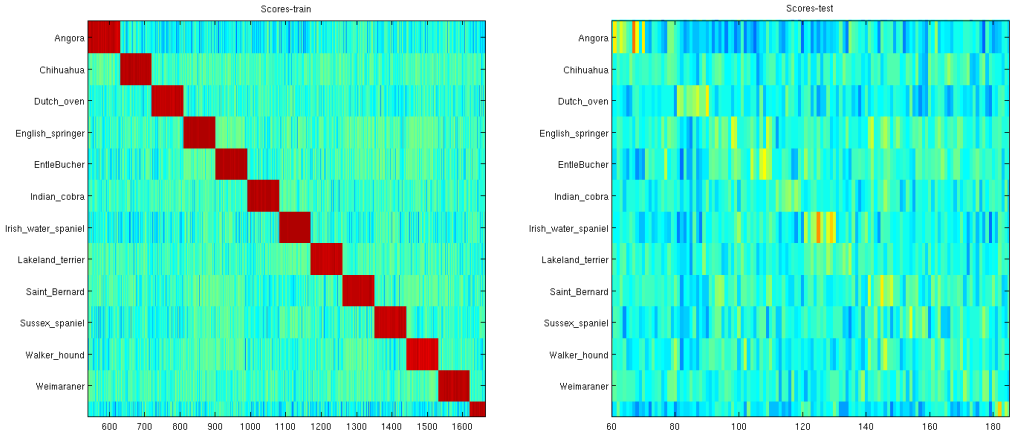

Figure 14.

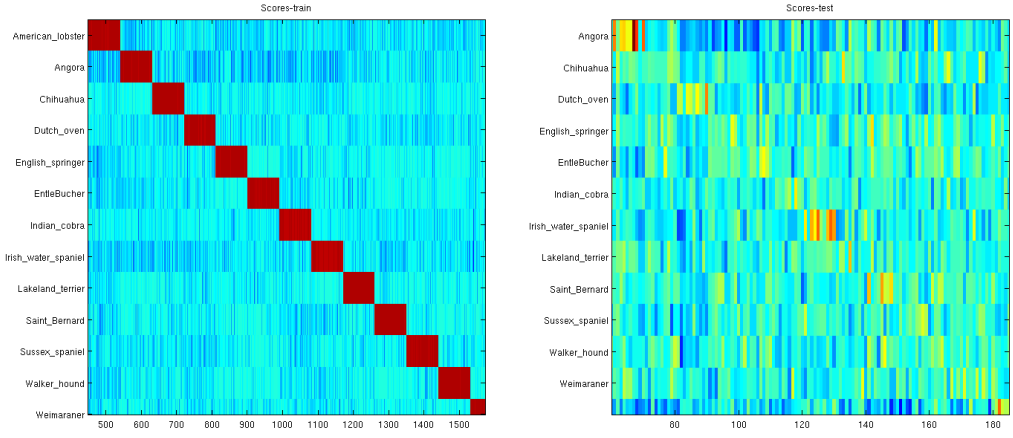

Figure 16.