Lab 8: PHOW Image Classification

Santiago Enrique Cortés
Bárbara & Frick
Av. Calle 82 # 10-33. Bogotá, Colombia
santiago.cortes@barbara.net.co

Juan Felipe Cerón Uribe
Universidad de los Andes
Cra. 1# 18A-12. Bogotá, Colombia
jf.ceron10@uniandes.edu.co

1. Introduction

In this exercise, we experimented with image classification in the Caltech 101 and a subset of the ImageNet image databases. In both of these, we applied the PHOW strategy to obtain a bag-of-words style set of features for images in the train and test sets. We then trained SVM classifiers on these features to attempt to classify the test images by class.

2. Dataset

The first dataset for our experiments was the Caltech 101 dataset. It contains from 40 to 800 images for each of 101 simple object categories. The images are quite minimalistic; they display the object clearly in the middle of the scene, it takes up most of the space and there is minimal presence of other objects. They do, however, show different instances of each category with varying rotations, illumination, depth and background. Another challenge is the presence of pictures as well as drawings and digital images.

Figure 1. A chair from the Caltech 101 dataset.

The ImageNet images appear slightly more difficult to classify. In them, the relevant category objects are often not in the middle of the scene, accompanied by objects that don't belong to the category and have more complicated backgrounds. Image 2 displays these characteristics except for a complicated background.

Figure 2. A coat from the ImageNet dataset.

3. Methodology

As mentioned before, our approach for this lab was to extract image features via the PHOW strategy and later classify them using SVMs.

3.1. Pyramid histogram of visual words (PHOW)

The PHOW strategy of feature extraction begins by the construction of a *visual vocabulary* form the training set. We follow a process similar to the construction of a texton dictionary:

1. Draw a grid on the image. Select a window size such that windows placed at each point on the grid will overlap.

2. Smooth the image proportionally to the grid step size (higher smoothing for a coarser grid) and estimate the intensity gradient at each pixel.

3. Divide the window around each of the grid points in 4, then compute the histogram of gradient directions at each subdivision based on 8 reference angles. The concatenation of the 4 histograms is a descriptor of each of the grid's points.

4. Repeat the last two steps at finer grids to obtain descriptors at different scales.

5. Cluster all of the descriptors by k-means. The k resulting centroids will be our *visual vocabulary*.

To obtain a bag of features for an image, follow the above steps except for the clustering. Instead, label each descriptor with its nearest centroid in the vocabulary. The PHOW description of the image will be the histogram of labels.

3.2. Classification

After building the vocabulary, we obtained the description of each image by the afore-mentioned process. We then trained a SVM on the training descriptions, and tested it on the test descriptions.

3.3. Hyperparameters

The most relevant hyperparameters for PHOW are the size of the vocabulary, the scales of the grids at which we will extract visual words, the bin size (alternatively the size of the feature-extraction window) and the amount of smoothing relative to the grid scale. Many of the sources we visited did not take the last one into account; we became aware of it while excecuting the provided script.

Because the concept of visual words is so abstract, we believe that the best way to approach a choice of hyperparameters should be based on cross-validation.

4. Comparison to SIFT

Although PHOW is based on SIFT, we believe the two algorithms to be best suited for different problems. SIFT is very good at finding particular instances of objects because it is both feature and rotation invariant, and because it is able to take into account the relative positions of keypoints from the object. It is not very good at finding objects belonging to a category because it tightly fits the characteristics of particular objects.

On the other hand, PHOW is best suited for classification tasks such as the one at hand because it learns more abstract visual words from several instances of each category. It also retains scale-invariance due to the consideration of grids and their accompanying window sizes at different scales. It is not well suited for instance detection because it isn't rotation-invariant on the local level. This is because keypoints (and thus visual words) aren't given an orientation, like in SIFT. However, some degree of orientation-invariance derives from the fact that we take histograms of visual words, disregarding their position and in particular the angles they form with respect to each other.

5. Experiments

We experimented with the described algorithm in subsets of both of the data sets. In both cases, we used 50 images per category as a training set and 35 as a test. All other parameters and hyper parameters were kept constant between data sets to make the different images and categories the actual basis of comparison.

6. Discussion

We obtained an ACA of 0.4966 for the Caltech data base and one of 0.2201 for the ImageNet data set. In the setup of both experiments, that is, with balanced data for every one of the classes this metric is a good estimator of model performance. The superior performance that the algorithm showed in the Caltech data set could be explained by the fact that this data set has fewer classes (and thus less room to make mistakes) and the provided code was tuned by its author for this data set. The images are also more complicated in the ImageNet data set, as discussed in section 2. In order to have a more detailed information about the results obtained, the F-measure for each category was computed. The result of those calculations are shown in the following histograms:

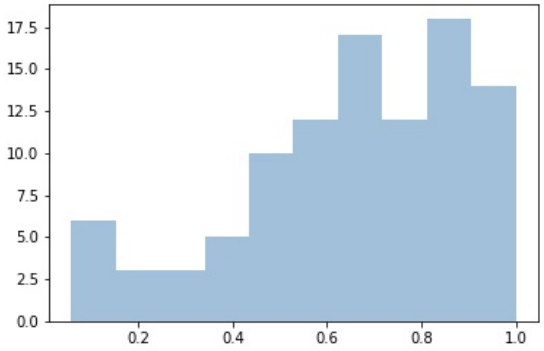


Figure 3. Histogram of F-measures of Caltech categories.

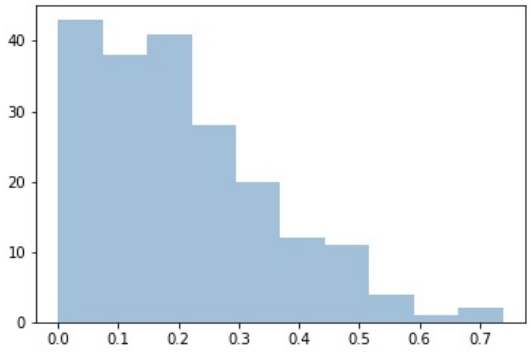


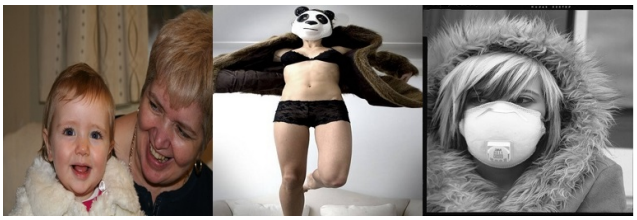
Figure 4. Histogram of F-measures of ImageNet200 categories.

In spite of its low ACA it can be appreciated that the PHOW algorithm is far superior on the CAltech Data set than on the imageNet200.

The classes with the best F-measure in the ImageNet200

data set were: whippet(0.73), zebra(0.60) and German short haired pointer. All these categories had quite homogeneous photos :

Given that the vocabulary is a parameter that is established equal for all the categories the categories whit several features in theirs photos are difficult to predict. An example of this phenomena can be seen in the categories with low F-measure, that also have bad precision and recall. Examples of the former are guacamole(0.095) and fur-coat(0.093).



The principal problems are clearly the variance of features between classes and the large number of classes in imageNet200. The principal challenge for the PHOW algorithm is the fixed number of vocabulary for all the classes. There are classes that require more features than others, also the SVM has even considering kernels have a a restriction in its boundary decisions. Even so, a way to improve the results using PHOW can be running the PHOW in batches of categories, to make fewer categories per run, another possible solution is to choose a Gaussian kernel for the SVM. Finally one possible suggestion to improve performance is to run several times per batch the algorithm and each pass change the vocabulary size in order to reduce the bias in the categories with high variance among its images.

References