Image Classification by PHOW

Alejandro Trujillo
Universidad de los Andes
Bogotá, Colombia
af.trujilloa@uniandes.edu.co

Santiago Martínez
Universidad de los Andes
Bogotá, Colombia
s.martinez1@uniandes.edu.co

Abstract

The task of image classification has been one of the main problems of computer vision. Multiple Datasets have come with time trying to generalize the problem so everyone can work on it with the same images and annotations. ImageNet is one of the newest datasets trying to approach the problem, having millions of annotated images and tens of thousands categories, they changed significanly the problem proposed by older datasets like Caltech101, which was one of the first big classification dataset. In this laboratory we will use a PHOW technique to generate descriptors of images and SVMs to classify them in the two aforementioned datasets. The ACAs obtained in the test set for Caltech101 and ImageNet were 71,0294 and 23,33 respectively, after tweaking multiple parameters for optimization. The huge difference in this numbers can be explained due to the difficulty of both datasets, as one is much newer than the other one, it approaches more realistic problems than the other one.

1. Introduction

The objective of recognition consists of acquiring semantic information directly from the image and classify it in a semantic category. The applications of recognition vary depending on the precision level needed to solve the problem. Some branches of the problem of recognition are:

- Image classification
- Object detection
- Object identification
- semantic segmentation

between others.
The main challenges when it comes to the task of recognition include; scale, occlusion, variations in point of view, deformation among others. Here, we'll use a PHOW approach in order to get a descriptor of the images and train multiple SVMs to classify them.

2. Methodology

2.1. Dataset Description

In this laboratory, two datasets where used in order to compare the performance of the PHOW algorithm over datasets with different difficulties. These were the *caltech101*, which is one of the first datasets disposed for the purpose of classification, and a small version of the *imagenet* dataset called *imagenet200*. The datasets were composed of 101 and 200 categories respectively. The difficulty of the database also varies, as imagenet is much newer, it represents a bigger challange than caltech101, even as its tiny version. The Caltech-101 datase consists of images distributed in 101 different semantic categories and Backround. each category has between 40 and 800 images. The image size isn't constant but it's close to 300x200. All images in the dataset are annotated. The objects in this dataset are mostly centered and the pose they present is also mostly stereotypical. This implies that the challenges of recognition in this dataset include simple variations among each class. [1] ImageNet is a image dataset which is structured according *WordNet* hierarchy, which is a lexical database of English words grouped in "synsets" (synonym sets). These are cognitive synonyms that represent different concepts. There are aproximately 100 000 synsets in ImageNet, However due to the difficulty of the task [2], we'll use a much smaller version of this dataset (ImageNet200), where instead one can fin 200 categories out of the original dataset. This dataset represents a more difficult task than the last one, as the images aren't in such stereotypical poses as those in caltech 101.

2.2. PHOW

Pyramid Histograms of Visual Words (PHOW) is an algorithm with which on can generate a shape descriptor for a given image. Firstly, it's necessary to compute the dense variant of the scale invariant Feature Transform (SIFT). This procedure consists of extracting features from patches of the image. In our case, the patches were divided into 4x4 structures, and 8 different orientations were used for the gradient. This means that the representation space only takes into account shape information until now. Afterwards, K-means is applied on this space in order to compute the ”visual dictionary”, the centroids calculated by K-means represent the words in this dictionary. After this, the histogram is calculated. The pyramid of PHOW consists of dividing the image into different patches and applying the aforementioned process to each patch, saving the information for each patch. This means that the feature space now also includes spatial information.

The difference between PHOW and SIFT is that the first one uses the dense variant of the latter, and uses different resolutions of the image. PHOW is a way to attack the problem of SIFT ignoring spatial information. This implies that PHOW is scale invariant as it applies SIFT in multiple levels.

Although The idea of PHOW sounds similar to that of textons, they differ in multiple manners. For instance, the scale used by the textons is much smaller, as it's designed to detect textures in an image, while PHOW tries to describe shape. Also, The textons filter bank approach is much more flexible than the oriented histogram one, as you can find very varied types of structures, This also makes textons much more expensive computationally compared to PHOW.

2.3. SVM

After obtaining the image descriptor using PHOW, multiple SVM classifiers were trained (one per category) in order to classify the images in their respective class.

2.4. Algorithm

The vl_feat library for Matlab was used in order to get the expected results, along with a code developed by Andrea Vedaldi, which uses PHOW and SVMs to classify images in the Caltech101 dataset. This code was also modified in order to use the ImageNet200 dataset in the same manner. The SVMs were only trained in the train set, and evaluated in the test set in order to analyze the effects of varying the hyperparameters of the method. The approach to evaluate these hyperparameters consisted of fixing all of them except one and tweeking it until getting the best performance, then move to another one and start over. The Hyperparameters varied in this paper were: The number of train images, the number of words in the dictionary, the size of the spatial partitioning and the value of the constant C in the SVM.

2.5. Evaluation Method

The evaluation method used for this problem was computing the confussion matrix and sum its normalized diagonal, which is the accuracy of the method (ACA).

3. Results

3.1. Effect of the number of categories

In order to visualize what the effect of the number of categories on the performance of the algorithm, the code was run using both datasets and different amount of categories.

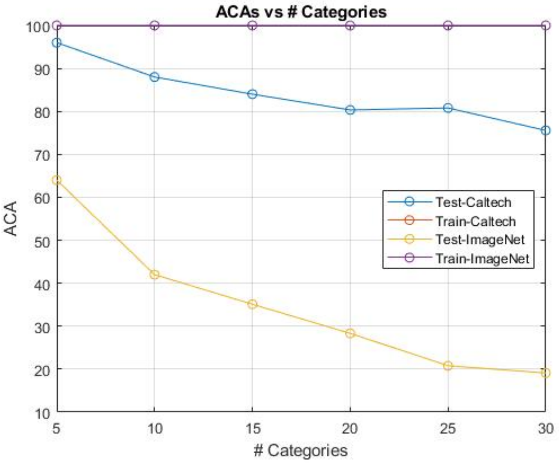

Figure 1. ACA vs # Categories

In figure 1, one can clearly see how the amount of categories affects the performance, as the more categories there is, the worse the algorithm performs. This is due to the Classification SVM being required to differentiate between more classes. Moreover, it's also evident the difference in difficulty overall between the datasets, given the performance of the algorithm over the same amount of categories, and how the performance deterioration given more categories is much higher in the ImageNet dataset.

3.2. Experiments

Firstly, the amount of train and test images were changed while the rest of the parameters remained constant. The default parameters given by the algorithm can be seen in table 1.

Hyper-Parameter	Deafault Value
Train & Test set size	15
Words in Dictionary	600
Spatial Partitioning (X & Y)	[2 4]
Step PHOW	3
C parameter (SVM)	10

Table 1. Default parameters for the algorithm

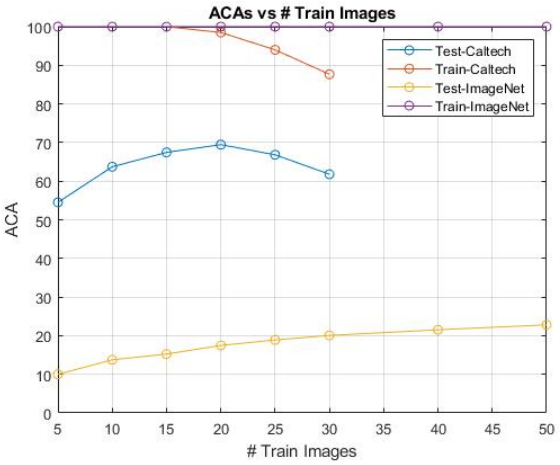

Figure 2. ACA vs # Train & Test Images

In figure 2, the number of Train and Test (always taken as the same) images was varied while the rest of the parameters remained constant. In case of the caltech101 dataset, a maximum is reached in 20 images, while in the ImageNet dataset a maximum ACA value was obtained with 50 images and due to lack of computational power we couldn't try higher numbers. Although the graph behaves as if it could grow more, we will use this value as the maximum. The amount of train and test images is fixed to the maxima in ACA obtained before.

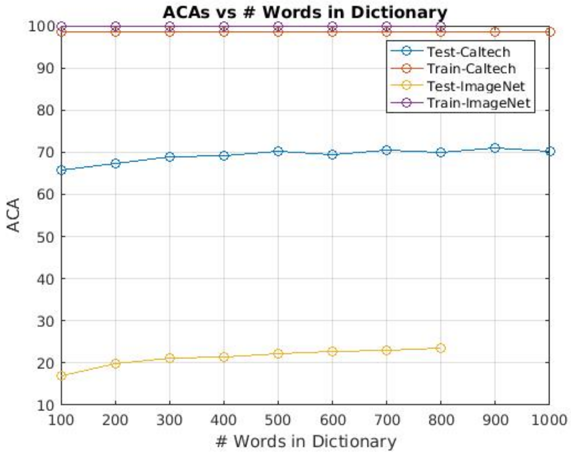

Figure 3. ACA vs # Words in the bag

While tweaking the mount of visual words used, we realized the more words we gave the algorithm in the case of ImageNet, the better it performed, However the computational cost didn't seem to be worth the slight better performance, so we decided to use a value of 700 words. In the case of Caltech, the best performances was yielded by 900 visual words. Just like textons, the amount of words has a rather unpredictable effect on the performance, as it uses K-means in order to cluster the visual words in the train set. It is highly dependant on the quality of the train set. However There is a minimum amount of words for the algorithm to work properly, that's why at the beginning of the curve the performance improves considerably and after that it's close to constant.

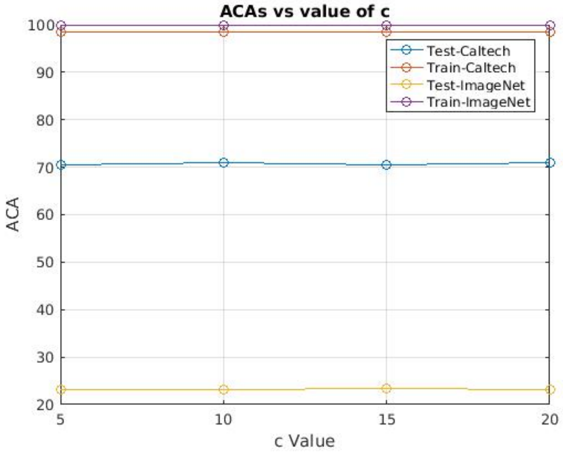

Figure 4. ACA vs c value of the SVM

In figure 4, one can see the effect of the c parameter of the SVM classifier on the performance of the algorithm. It's evident that the performance didn't change much while tweaking this parameter, this is perhaps due to all of them being similar values. Maybe if it was tried with much larger values the effect could've been seen. However not knowing exactly how the objects are being represented in the representation space, it's rather difficult to get this value to change the performance significantly.

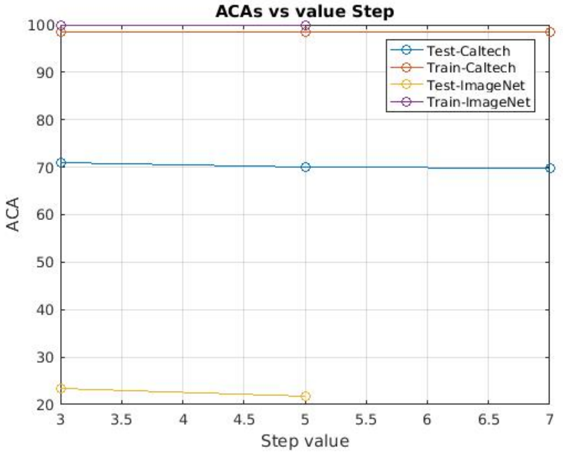

Figure 5. ACA vs Step value of PHOW

The step didn't affect significantly the performance of the algorithm as one can see in figure 5. It reduced the performance of the algorithm significantly in the case of ImageNet when changed from 3 to 5. This is due to the PHOW extracting less information as the step gets larger. for this reason it was decided to keep the step on 3.

3.3. Final values for Hyperparameters

Hyper-Parameter	Optimal Value (Caltech101)	Optimal Value (ImageNet)
Train & Test set size	20	50
Words in Dictionary	900	700
Spatial Partitioning (X & Y)	[2 4]	[2 4]
Step PHOW	3	3
C parameter (SVM)	10	15

Table 2. Default parameters for the algorithm

In table 2 appear the hyper-parameters chosen for each dataset.THE ACA for Caltech101 and ImageNet200 were 71,0294 and 23,33 respectively. Such large different is due to the difference in difficulty in bot problems.

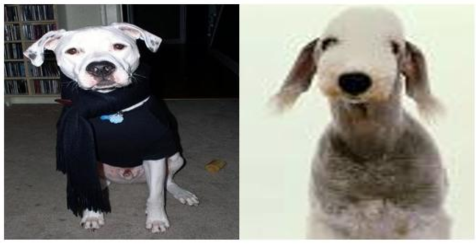

Figure 6. Example imageNet200

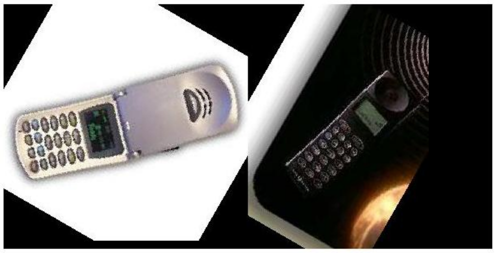

Figure 7. Example Caltech101

One of the least succesful categories for our algorithm in imageNet was the American Staffordshire terrier. Looking at figure 6 one can see how difficult the task was, on the left we have an image from the aforementioned category, while on the right hand we have a completely different category, however one can see the huge similarity in shape. ImagNet200 has multiple classes that are dog's breeds apart from this two. On the other hand one of the best categories was the cellphones category of caltech. However, one can see in figure 7 a huge bias in shape in this images, as they're all rotated, leaving a dark space on the edges of the image, which simplifiesthe problem a lot while using shape.

4. Conclusions

From the results obtained we can conclude that the ImageNet dataset is by far more difficult than the caltech101 dataset. This is due of it including most of not all challenges of recognition. Being PHOW an algorithm that describes shape only, it results convenient for the caltech101 dataset, however, shape only isn't enough for it to be adequate for the ImageNet dataset. One way to improve the result would be to enhance the descriptor by adding color information, local color difference and texture to it. Moreover, as the problem is not a binary classification one, maybe an SVM isn't the best option. Perhaps trying a Random Forest classifier or a KNN (K-Nearest Neighbors) one could get better results as these classifiers are better designed for multi-class problems.

5. References

[1]L. Fei-Fei, R. Fergus and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories", Computer Vision and Image Understanding, vol. 106, no. 1, pp. 59-70, 2007. Available: 10.1016/j.cviu.2005.09.012.
[2]"ImageNet", Image-net.org, 2019. [Online]. Available: http://image-net.org/about-overview. [Accessed: 24- Mar-2019].