PHOW Image Classification

Mauricio Neira
Universidad de los Andes
Cra 1 Nº 18A - 12, Bogotá - Colombia
m.neira10@uniandes.edu.co

Daniel Rodriguez
Universidad de los Andes
Cra 1 Nº 18A - 12, Bogotá - Colombia
da.rodriguez1253@uniandes.edu.co

Abstract

We use the PHOw algorithm to classify images in the Caltech 101 and the Imagenet dataset. We find that the best parameter configuration of PHOW was given with a $ws = 5$ and a $rss = 0.5$ as they are the values that extract the most rich information in the feature space. Likely because the images in the search have the target classes in a part of the image and not the whole. Additionlly we find that fine grain classification is significantly harder for the algorithm and thus performs poorly in these situations.

1. Introduction

Image classification is one of the most important ongoing problems in modern computer vision. There are a myriad of applications in the industry and science ranging from astronomical object classification to quality control.

In this study we evaluate PHOW as a method for image classification on two datasets: Caltech 101 and Imagenet.

1.1. PHOW

Pyramid histograms of visual words or PHOW is a method used to classiy images. The method in itself is very similar to SIFT (Scale Invariant Feature Transform). The basic method involves transferring the image into a feature space where a classifier like an SVM can then predict the class of an image based on its feature representation given that a good enough training set is given.

The strength of this method relies on the way it creates the feature space. It generates a window with 4×4 sub-windows. Every $1/16$th of a window has a size $ws \times ws$ (ws is the window size of every $1/16$th sub-window and will be referenced as window size from now on). On every sub-window, 8 directional gradients are calculated giving a total of $8 \times 16 = 128$ features for every window. In this $\mathbb{R}^{128}$ space, we run a k-means algorithm to get k clusters. Each pixel is then mapped to the nearest cluster so that each pixel has a value of $k \in \{1, ..., k\}$. A histogram for the image is calculated based in these values so that each image is represented in a $\mathbb{R}^k$ space. A classifier is trained on this space to predict future images.

2. Databases

Two main databases were used in this study Caltech 101 and Imagenet.

2.1. Caltech 101

Caltech 101 is a databse comprised of a total of 101 categories. The amount of images per category vary from 40 to 800. The size of each image is around 300×200. Almost all of the images are natural images and man made objects with categories inluding joshua trees, water lilys, ceiling fans etc. Some are shown in figure 1.

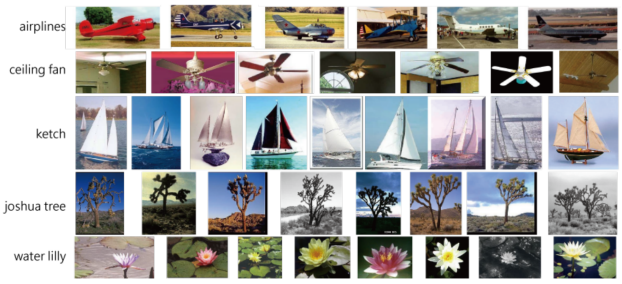

Figure 1. Example images from the Caltech 101 database. Courtesy of [2].

2.2. Imagenet

Imagenet is a database where images are taken from the WordNet hierarchy. Each word has approximately 500 images. Some images are shown in figure 2.

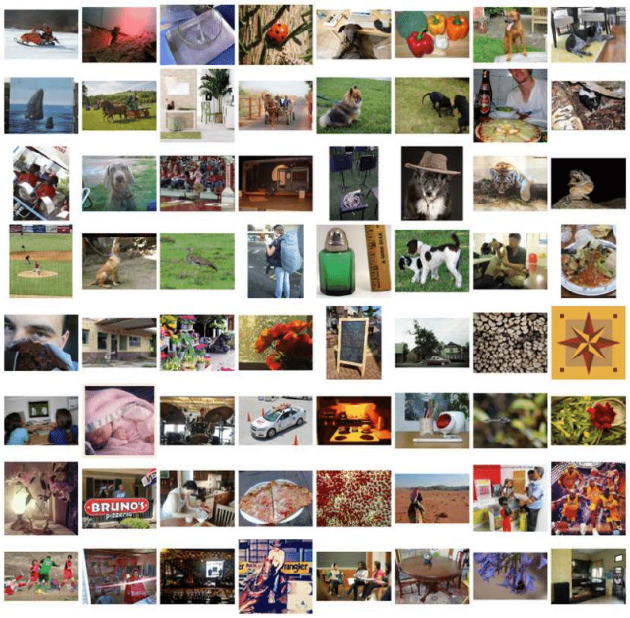

Figure 2. Example images from the Imagenet database. Courtesy of [1].

3. Approach

3.1. Implementation Details

We used the vl_feat library[1] and used the default value of k. All other parameters were studied.

4. Experiments and discussion

4.1. PHOW on Imagenet

The first experiment was to determine the optimal window size and step size for the algorithm. For that purpose, we fixed every other parameter in place in the following way:

- SVM cost $C = 10$
- 15 images per class on the training set
- 20 classes

We sweeped the window size ws for the values in $\{5, 10, 15, 20, 40\}$ and varied the step size *relative to the window size* rss in the following proportions $\{0.15, 0.3, 0.5, 1\}$. We calculated the absolute step size $asts$ in the following way: $asts = \lfloor ws \cdot rss \rfloor$. The results can be seen in figure 3.

The best performing algorithm was found with $ws = 5$ and $rss = 0.5$. These results are surprising specially considering that the ACA *increases* as the window size decreases. This could mean that the feature representation finds richer content in local sections of the image. This makes sense considering that many of the objects in the imagenet dataset do not comprise the whole image but only a part of it.

With the best performing algorithm, we carried out additional experiments to see how other parameters played a role in the classification. The first experiment consisted in varying the number of classes to classify. In figure 4, it is evident that the performance of the algorithm decreases as the number of classes increases. This is natural considering that classifying more images involves classifying images that are more closely related. That is to say, as the number of images increases, the problem resembles a fine grained classification problem which is notoriously difficult.

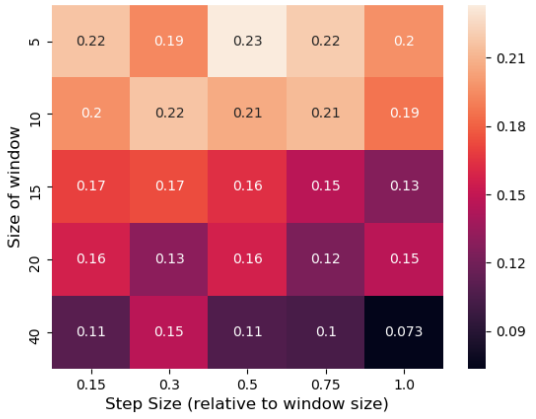

Figure 3. Heatmap of the ACA on 20 classes in Imagenet varying the step size and the square length of each $1/16$th of the window.

Additionally, figure 5 demonstrates that the performance of the algorithm increases with an increase in the number of images per class in the training set. This is easily explained as with a greater amount of examples, the SVM can draw a hyper-plane with greater confidence in each class. In general, machine learning algorithms perform better with more training examples.

Running the algorithm with parameters $ws = 5$, $rss = 0.5$, $C = 10$ and using all of the training images available on all of the 200 classes of Imagenet yields an ACA of 11.4%. Altough this ACA is far from desirable, considering that guessing an answer constitutes $1/200 = 0.5\%$ ACA, we can affirm with confidence that PHOW is learning to classify the images.

Due to the size of the confusion matrix, it is not included in this section but it can be found in the suplementary section at the end of the document. The behaviour of some classifications are worth noting though. For instance, the

[1]See `http://www.vlfeat.org/index.html` for more information.

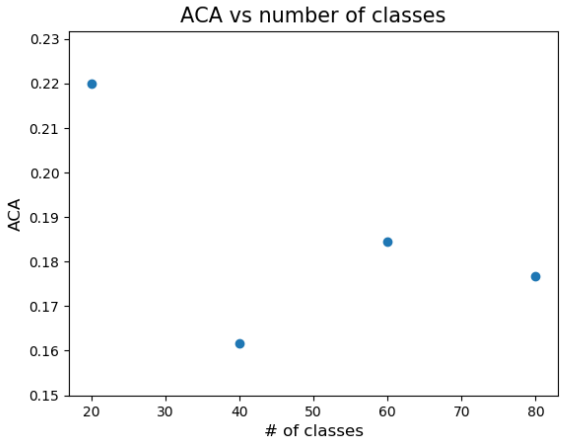

Figure 4. ACA as a function of the number of classes on the Imagenet dataset. The algorithm was run with a stepsize of 1, a window size of 5 and 15 images per class on the training set.

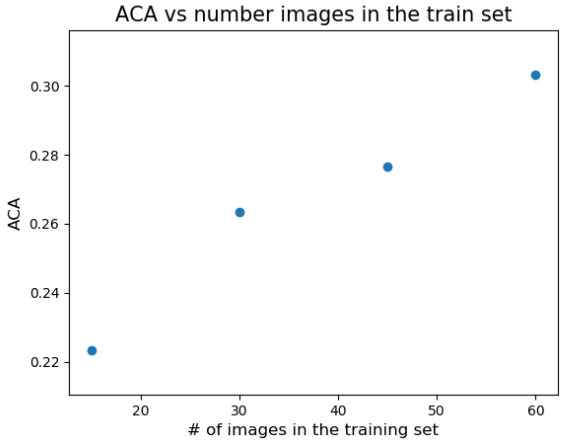

Figure 5. ACA as a function of the number of images in the training set on the Imagenet dataset. The algorithm was run with a stepsize of 1, a window size of 5 on a total of 20 classes.

submarine class and the suit class are classified particulary well. This comes as no surprise since these classes have no similar classes. The class representing silky terriers on the other hand are very often confused with howler monkeys. This makes sense as those two classes are similar. They represent animals of similar texture and color. Something similar occurs with dowtichers (a type of bird). The are very often confused with spoonbills (another type of bird). This algorithm can distinguis classes that are *very different*. Like submarines and suits. However, it underperdorms significantly in fine grain differences.

4.2. PHOW on Caltech 101

The total ACA on the caltech 101 database was 56.99% significantly higher that the imagenet ACA. This reflects the dificulty of the dataset relative to imagenet. The reduced amount of classes and the difference between each class makes the classification problem a lot easier.

5. Conclusions

Based on the results we obtained we conclude that the performance of PHOW decreases as num classes increases since more classes implies that the algorithm has to distinguish between classes that are more similar. This result is confirmed by varying the number of classes in the imagenet dataset and when comparing the ACA between the Caltech 101 and the Imagenet database.

Secondly, the performance of PHOW increases proportional to the amount of training instances available. This result is seen accross all areas of machine learning where a larger dataset enables the algorithm to make a more accurate generalization of the data.

We found that the best parameter configuration of PHOW was given with a $ws = 5$ and a $rss = 0.5$ as they are the values that extract the most rich information in the feature space. Likely because the images in the search have the target classes in a part of the image and not the whole.

References

[1] C. Chen, Y. Ren, and C.-C. J. Kuo. Global-attributes assisted outdoor scene geometric labeling. In *Big Visual Data Analysis*, pages 93–120. Springer, 2016.

[2] S.-C. Wang and Y.-C. F. Wang. A multi-scale learning framework for visual categorization. In *Asian Conference on Computer Vision*, pages 310–322. Springer, 2010.

Suplementary material

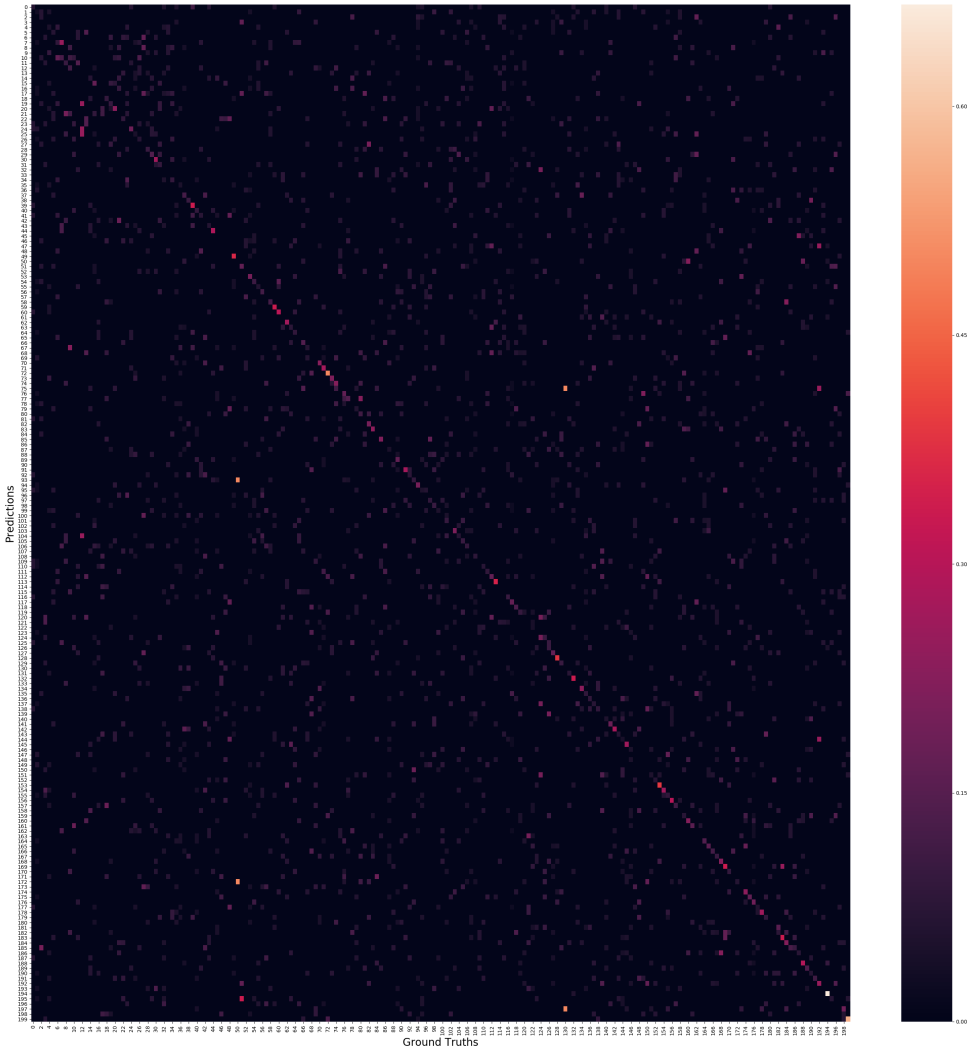

Figure 6. Confusion matrix for the 200 classes in Imagenet.