PHOW Image Classification

Maria Camila Escobar
Universidad de los Andes
Bogotá D.C, Colombia
mc.escobar11@uniandes.edu.co

Laura Gongas
Universidad de los Andes
Bogotá D.C, Colombia
l.gongas10@uniandes.edu.co

***Abstract*-ImageNet is a dataset that changed significantly the problem of image classification with respect to older datasets such as Caltech 101. Unlike Caltech 101, ImageNet approaches the majority of the challenges of recognition which are view point variation, illumination, occlusion, scale, deformation and background clutter. This paper explores the PHOW strategy for object recognition, specifically image classification, in the datasets Caltech 101 and ImageNet. The best ACAs obtained in the train sets were 69.74 and 20.97 for Caltech 101 and ImageNet respectively. The evaluation of the test set of ImageNet with optimal parameters resulted in an ACA of 20.925.**

1. Introduction

The objective of recognition problems in computer vision is to extract information from the world through given images. Recognition applications vary on precision levels of the problem. For instance, image classification, object detection, object identification, semantic segmentation and 3D pose estimation, are variants from recognition problems. For image understanding, a combination of the previous variants are necessary [6]. Challenges of recognition include: view point variation, illumination, occlusion, scale, deformation and background clutter.

The Caltech 101 dataset are pictures of objects from 102 different categories. Each category has between 40 and 800 images, but most of them contain only 50. The size of each image is approximately 300x200 pixels. There are annotations for all of the images that consist of a bounding box of the object and a traced silhouette of the objects. The majority of images have no clutter, most of the objects are centered in the image and they're presented in a stereotypical pose [1]. In other words, this dataset only addresses the challenges of intra-class appearance variation and in a smaller extent, illumination.

ImageNet is an image dataset structured according to WordNet hierarchy. WordNet is a lexical database of English in which nouns, verbs, adjectives and adverbs are grouped into "synonym sets" or "synset". Groups of synsets are cognitive synonyms that represent different concepts. There are more than 100,000 synsets in WordNet and the majority of them are nouns [2]. A very small subset of this dataset was used for this paper (200 categories of objects with 100 images each). Images contains clutter, occlusion, and variations in lighting and pose [10].

Image classification with Pyramid histograms of visual words (PHOW) begins by computing dense Scale Invariant Feature Transform (SIFT) on training images. The main idea of SIFT is to extract features from patches of the image. To do so, each patch is divided into 4x4 cells. The orientation of the gradient is calculated for 8 different orientations in each cell. Building a representation space that takes into account shape. Then, a visual dictionary is constructed applying k-means on the SIFT space. The visual words that constitute the dictionary are the resulting centroids of k-means. Afterwards, a pyramid of visual word histograms is computed for each image. The spatial pyramid partitions each pyramid level in cells and calculates histograms for each quadrant. Finally, binary SVMs are trained for each category. Subsequently, in the test step, a pyramid of visual word histograms is computed for each test image. All SVMs are evaluated and the image is classified based on the highest confidence [7].

The difference between SIFT and PHOW is that the latter calculates dense SIFT at different resolutions. PHOW was suggested as a way to address the problem of histograms ignoring spatial information in SIFT. Also, PHOW is scale invariant because it employs SIFT, which is a scale invariant method to extract features. Due to the fact that PHOW considers histograms at different levels of a spatial pyramid, this model includes spatial information in a soft manner [9].

The objective of this paper is to explore the PHOW strategy for object recognition in two datasets: caltech 101 and ImageNet.

2. Materials and methods

The vl_feat library was used along with an algorithm developed by Andrea Vedaldi to classify images with PHOW. The datasets employed were caltech 101 and a small subset

of ImageNet. Experiments were performed in the train set of both datasets to analyze the effect of varying a hyperparameter while maintaining the other parameters fixed. We decided the which were the best parameter values based on ACA and time. The best paramater found in each experiment was fixed for following experiments. Finally, we obtained the hyperparameter configuration for optimal results in both datasets.

For the hyperparameter experiments we varied the number of categories to observe the change in results but all the experiments were performed taking into account all the classes of each dataset. Also, we varied the number of train and test images, the number of dictionary words, the C value for SVM, the spatial partitioning, and the number of trees for the KDTree quantizer.

Afterwards, we developed a function to classify the test set of ImageNet based on our optimal train model.

3. Results

The most relevant hyperparameters for the PHOW strategy are the number of dictionary words, spatial partitioning, sizes and step, which will be explained later. Other hyperparameters that we consider important are: the color of the image (rgb and hsv options) because it provides more information for the descriptor, the size of the Gaussian window used to smooth the image, the number of neighbors for the KD tree query, the number of trees for the KD tree quantizer, and the C parameter of SVM (the latter will be discussed later). The kernel used in SVM is important too because the default is chi squared but the intersection kernel could also be appropriate. Also, studies have shown that the generalized Gaussian kernel yields better results [6]. Additionally, the epsilon parameter of SVM might be an important parameter for the model because it determines the incorrect classification permitted.

The best set of hyperparameters for both datasets was chosen by performing experiments in which all the parameters were fixed except one. The value of the varied parameter that yielded the better result (in terms of ACA and time) remained fixed for further experiments. In all the tables, the ACA in bold represents the chosen value of the parameter.

Table 2 shows the ACA for caltech101 and ImageNet datasets with default parameters fixed (Table 1) but varying number of categories. In this case, we decided to perform posterior experiments with all the categories of each dataset (102 for Caltech 101 and 200 for ImageNet) even if the ACA worsened for larger number of categories. These results were expected because more classes mean there is a larger range of class-confusabilities [8].

Parameter	Value
Train set size	15 images
Test set size	15 images
# Categories	5
Dictionary words	300
SVM C	10
# Trees	1
Spatial partitioning	2x2
Size PHOW	7
Step PHOW	5

Table 1. Default parameters of Andrea Vedaldi algorithm to classify images with PHOW.

	ACA	
# Categories	Caltech 101	ImageNet
5	92	54.67
50	61.87	14.67
Max. in dataset	**58.24**	**10.57**

Table 2. Variation of number of categories for Caltech 101 and ImageNet datasets. Other parameters were fixed at default values.

Then, the number of images for train and test were varied for Caltech 101 and ImageNet (Tables 3 and 4 respectively). For both datasets, better results were obtained when the train set size increased. The result makes sense because the model has more information for posterior classification. However, the results were not consistent between datasets for the test set size. For Caltech 101, a smaller amount of test images yields a better result. If less images have to be classified then there is a smaller probability that the model will commit errors. It is important to note that the smaller number of test images was chosen for further experiments because it takes significantly less time for hyperparameter experiments.

	Train		Test	
# Images	ACA	Time (s)	ACA	Time (s)
15	58.24	87.47	**60.85**	97.35
20	**60.85**	97.35	59.31	116.25

Table 3. Train and test set size variation in the Caltech 101 database.

	Train		Test	
# Images	ACA	Time (s)	ACA	Time (s)
15	10.57	186.52	**12.63**	472.73
30	11.7	286.86	12.6	550.60
50	**12.63**	472.73	12.85	639.90

Table 4. Train and test set size variation in the ImageNet database.

The first relevant hyperparameter of the PHOW strategy that was tested was the number of words in the dictionary. The number of dictionary words is equivalent to the k parameter of k-means and therefore it must be determined experimentally. Based on literature, few words result in a small dictionary that is not representative, while many

words tend to overfitting and similar objects are divided into different clusters [6]. Table 5 exposes the results for varying amount of words in both datasets. It is evident that a small amount of words does not represent well the data. However, large amount of words experiments were not carried out because they were too time expensive.

	Caltech 101		ImageNet	
# Words	ACA	Time (s)	ACA	Time (s)
100	53.09	95.41	9.8	424.53
300	60.85	97.35	12.63	472.73
500	60.98	127.33	**13.83**	547.52
700	61.57	136.82	13.87	585.78
1000	**64.77**	154.02	-	-

Table 5. Variation of the number of dictionary words in PHOW for Caltech 101 and ImageNet datasets.

Then, the optimal C parameter of SVM was chosen for each dataset based on the results of Table 6. Even though the ACA does not vary a lot for different C values, it is a very important parameter because it determines the importance of the two terms in the soft margin formulation of SVM. This formulation includes a term for maximizing the margin between the categories and another term for misclassification through a slack variable. The C parameter determines the weight of the latter, in other words, how much we want to avoid misclassifying each training example [11] [7].

Large values of C result in a smaller training error. However, if it is too big then the classifier will try to classify everything correctly and it might loose generalization. This explains why the optimal C value for both datasets is not 1000 (Table 6) [5]. In contrast, if the C value is small, the classifier will look for a large margin separating the hyperplane even if this implies misclassification. This will happen even if the data is linearly separable [11]. The ACA for a small value of C yields the worst result for the Caltech 101 dataset, which is consistent with literature. Nonetheless, the best ACA for ImageNet was obtained with the smallest C value. This might be caused because the data is more linearly separable in this dataset.

	Caltech 101		ImageNet	
SVM C	ACA	Time (s)	ACA	Time (s)
0.1	62.81	138.2911	**14.33**	434.9984
10	**64.77**	154.0236	13.83	547.5234
1000	64.38	151.183	13.23	563.192

Table 6. Variation of the C parameter of SVM in PHOW for Caltech 101 and ImageNet datasets.

Table 7 shows a significant increase in results for a greater spatial partitioning which makes sense because the descriptor has more information to represent each image.

	Caltech 101		ImageNet	
Spatial partitioning	ACA	Time (s)	ACA	Time (s)
2x2	64.77	154.02	14.33	435.00
4x4	**65.56**	240.13	**16.33**	857.96

Table 7. Variation of the spatial partitioning in PHOW for Caltech 101 and ImageNet datasets.

The PHOW algorithm proposed by Vedaldi employs kd-tree as a data structure to solve a large scale nearest neighbor query. An important parameter for this method is the number of trees because it improves the effectiveness of the representation in high dimensions[3]. This is consistent with the results in Table 8 which show an increase of ACA as the number of trees got higher.

	Caltech 101		ImageNet	
# Trees	ACA	Time (s)	ACA	Time (s)
1	65.56	240.13	16.33	857.96
10	**67.39**	265.63	**17**	975.19

Table 8. Variation of the number of trees in the KD Tree quantizer for Caltech 101 and ImageNet datasets.

The sizes and step of PHOW are other important hyperparameters of the method. The sizes represents the scales at which the dense SIFT features are extracted. The step refers to the step of the grid at which the dense SIFT features are extracted [4].

Table 9 shows the results of the experiments varying the sizes of PHOW. First, only one level of the pyramid was considered and then, two and three levels. When more scales are considered, the descriptor is richer because it has more information and therefore better results are obtained. However, the optimal sizes of the windows are bigger in Caltech 101 than in ImageNet. When the scale values are bigger, thicker shapes are detected and when the scale values are smaller, more detailed shapes are obtained. In this manner, the difference of optimal sizes between the datasets is probably due to the different characteristics of images.

Similarly, a smaller step in PHOW constructs a more rich descriptor and this results in higher ACA (Table 10). A small step means that the dense SIFT features will be calculated more close to each other which, again, includes more information in the descriptor.

	Caltech 101		ImageNet	
Size PHOW	ACA	Time (s)	ACA	Time (s)
7	67.39	265.63	17.00	975.19
6, 8	66.14	333.67	18.13	1512.61
6, 8, 10	**68.30**	449.00	18.33	1885.27
4, 6, 8	66.80	449.08	**19.70**	1900.52

Table 9. Variation of the sizes of PHOW for Caltech 101 and ImageNet datasets.

	Caltech 101		ImageNet	
Step PHOW	ACA	Time (s)	ACA	Time (s)
3	**69.74**	1146.73	**20.97**	3378.96
5	68.69	618.99	19.70	1900.52
7	68.30	449.00	18.90	1384.50

Table 10. Variation of the step of PHOW for Caltech 101 and ImageNet datasets.

Table 11 summarizes the optimal hyperparameters found from the previous experiments for both datasets.

	Caltech 101	ImageNet
Train set size	20	50
Test set size	15	15
# Categories	102	200
Dictionary words	1000	500
SVM C	10	0.1
# Trees	10	10
Spatial partitioning	4x4	4x4
Size PHOW	6, 8, 10	4, 6, 8
Step PHOW	3	3

Table 11. Optimal parameters for the PHOW algorithm for the Caltech 101 and ImageNet datasets.

The best ACAs obtained evaluating in the train set were 69.74 and 20.97 for Caltech 101 and ImageNet respectively. The algorithm performed significantly better for the Caltech 101 than for ImageNet. The large variation of results is caused because the Caltech 101 dataset has images with the majority of objects centered in the image in a stereotypical pose and there is no background clutter. In contrast, ImageNet provides various objects per image, cluttered background, occlusion and variation in illumination and pose, making the dataset more difficult for classification than Caltech 101.

Additionally, we evaluated the performance of the optimal model for ImageNet and we obtained an ACA of 20.925. This value is similar to the one obtained for the train set (20.97). In other words, images are either very similar in the train and test sets and therefore the model works well for both or the model is generalized.

Based on confusion matrices from the hyperparameter experiments, the classes that seemed to be easier in ImageNet are web site and coral fungus. The web site category doesn't have any clutter, changes of illumination or occlusion, and the view point is minimally varied. This makes the category easier to classify thanks to its nature. However, about 40 categories didn't have any correct classification, which is 20% of the classes in the dataset. One of this difficult categories was chihuahua, which includes a very big variation inside the class and many different viewpoints (Figure 1). Even though the majority of the images don't have any change of illumination, clutter or occlusion, the previously mentioned challenges make the classification significantly difficult.

Figure 1. Images from the chihuahua category that qualitatively exemplify the difficulty of classification.

4. Conclusions

In conclusion, the ImageNet dataset is much more challenging than Caltech 101 because it includes all the challenges of recognition (view point variation, illumination, occlusion, scale, deformation and background clutter). Therefore, ImageNet is a more realistic problem of recognition. Based on our ImageNet results (ACA of 20.925 in test set) compared with Caltech 101 (ACA of 69.74 in train set), the PHOW algorithm that performs really well on Caltech 101 is not adequate for ImageNet.

To improve results we recommend adding texture and color information to the descriptor, providing a richer representation of the images. Also, if the categories were divided in subcategories based on different poses of the object, the model would learn easier. Nonetheless, the last recommendation would require a modification to the dataset and a significant increase of categories. Finally, due to the fact that the problem is multi-category, SVM might not be the most adequate classifier so we recommend to employ a classifier that is not binary.

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. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004.

[2] *ImageNet*.

[3] Vlfeat tutorials - kd-trees and forests. *Vlfeat.org*.

[4] Vlfeat documentation - vl_phow. *Vlfeat.org*, 2018.

[5] I. Ahmed. In support vector machines (svm) how can we adjust the parameter c? why is this parameter used? *ResearchGate*, 2012.

[6] P. Arbelaez. Recognition 01. 2018.

[7] P. Arbelaez. Recognition 02. 2018.

[8] M. Gupta, S. Bengio, and J. Weston. Training highly multiclass classifiers. *Journal of Machine Learning Research*, 15:1461–1492, 2014.

[9] S. Mahdi and K. Razavi. What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. *Arxiv.org*, 2018.

[10] P. Sermanet, A. Frome, and E. Real. Attention for fine-grained categorization. *Arxiv.org*, 2015.

[11] M. Shivers. What is the influence of c in svms with linear kernel? *StackExchange*, 2012.