Image classification using pyramid histograms of visual words

Daniel Viteri Noguera
Universidad de los Andes
d.viteri@uniandes.edu.co

1. Introduction

When addressing a computer vision classification problem, it is useful to extract visual features from images. These are called descriptor and are used to train models to solve this problem. The histograms of oriented gradients (HOG) are the most commonly used ones [1], and are the benchmark of more sophisticated descriptors as Scale Invariant Feature Transform (SIFT) and histograms of visual words (PHOW), under the idea of bag of words [2].

In order to obtain the SIFT descriptor, a a window is extracted from an image key point. The window, is divided into a 4X4 patch and inside every patch, HOG will be calculated for different angles (8 as default), Making it a scale invariant descriptor. This way, the SIFT descriptor is obtained and can be used to train a classification model to learn the vocabulary [2].

Similarly, PHOW is a dense SIFT descriptor, in which even more fine grids is placed in every patch of the original division. Taking this into account, PHOW would also be scale invariant descriptor. This work will evaluate the PHOW descriptor in the Caltech101 and ImageNet dataset [3].

The Caltech101 dataset, was collected by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato in 2003. It is made up of 101 different classes of small images (roughly 300 x 200 pixels). Every class have from 40 to 800 images (most of them have 50)[4]. This is now considered a solved database, and now the imagenet is the one being used for research. Caltech dataset include catalog images, that make the problem much simpler. Sample images from the dataset can be seen in figure 1.

Currently, the imagenet database is being collected. It is organized according to the WordNet hierarchy, which at the times has only nouns. Its images, are high resolution, and different sized [5]. The imagenet project pretends to have about 50 million annotated images by the time it is completed. Sample images from the dataset can be seen in figure **??**.

Figure 1. Sample images from the Caltech101 dataset.

Figure 2. Sample images from the ImageNet dataset.

2. Method

The PHOW approach considered in this work consists in a sliding window with certain step, which is divided into a grid to obtain the HOG (basically extracting the dense SIFT). Before this happens, the image is filtered to make it smoother and facilitating the task. Then, this descriptor is trained for each class using k-means, for further classification.

Three experiments were made, varying the hyperparameters of the function:

First, the number of words was examine to determine its effect in the accuracy of the algorithm. The other parame-

ters were set as default (Size= [4 6 8 10] and Step= 2). Also, the number of words went from 100 to 1000 with steps of 25.

Second, the window size was changes, setting the other parameter as default (Number of words= 600, Step=2). This parameter was changes from 2 to 25.

Finally, the step parameter was varied, setting the other parameters as default (Number of words= 600, Size= [4 6 8 10]). It was changed from 1 to 25.

Also, the hardest class from each dataset was extracted by obtaining the minimum value of the diagonal of the confusion matrix, and labeling it with its corresponding class.

3. Discussion

One of the parameters of the PHOW approach worked in this paper, is the size of the bag of words. According to literature, researchers tend to choose it empirically [6].

In figure 3, the response of the ACA to changes in the number of words in the Caltech dataset is shown. This shows that in a range from 100 to 1000, this parameter is not significantly relevant. This same behaviour is observed in figure 4, which is the same graph but obtained from the Imagenet dataset.

Given that this ranges was taken arbitrarily, there is not previous knowledge about the relevance of the visual words that the algorithm is considering. Therefore, what these results might prove, is that there are many redundant words that are being selected to train the model. Further work should be focused on determining the relevance of the number of words, which are basically the clusters of k-means.

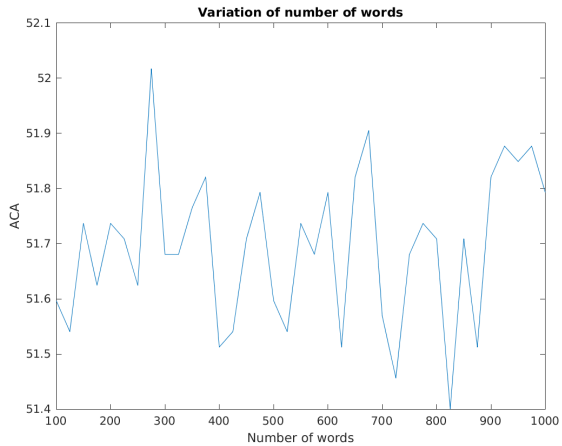

Figure 3. ACA when varying the number of words in Caltech.

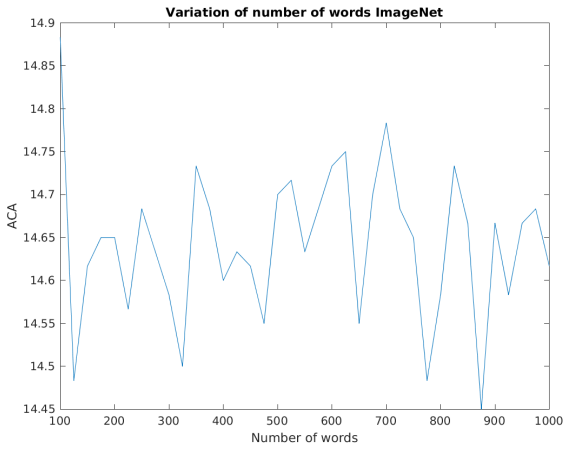

Figure 4. ACA when varying the number of words in ImageNet.

Another hyper parameter is the window size, or the scale at which the PHOW descriptor is being extracted. Figure 5 shows the ACA response to this variation in the Caltech dataset. Similarly, figure 6 displays this same analysis from the ImageNet dataset. Both, exhibit a decrease in the algorithm accuracy when the window size increases in a range of 2 to 25 pixels.

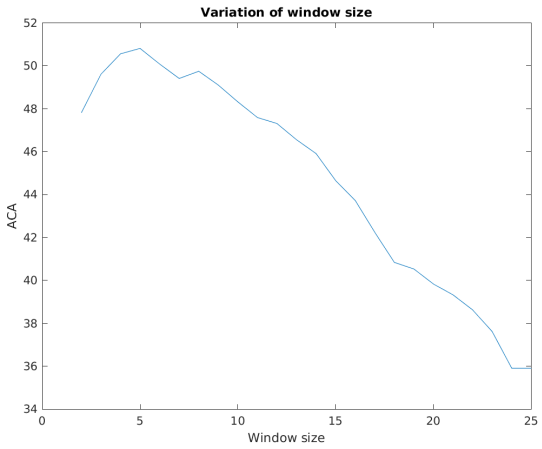

Figure 5. ACA when varying the window size in Caltech.

This behaviour, is explained due to two factors. The first one, is the fact that when the size of the window increases, small details from the image might be kept out from the descriptor. The second one, is that this parameter was set using a number, not a vector. This, means that the information was being extracted using only one scale, instead of various, affecting the outcome of this analysis and the ACA. However, it can be concluded that the window size

affects significantly the accuracy of the algorithm, and must be set taking into account the image size and the nature of the objects.

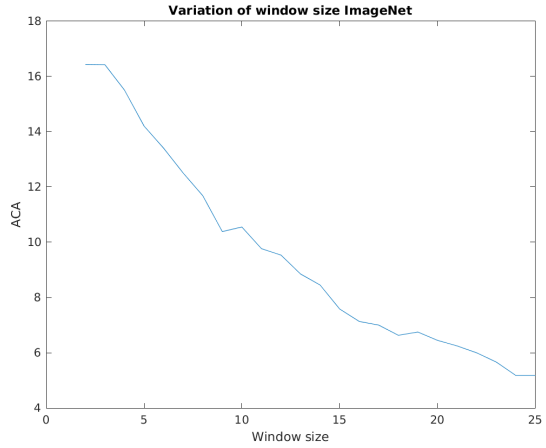

Figure 6. ACA when varying the window size in ImageNet.

Finally, the variation in the step hyperparameter was made to see its response in the ACA of the algorithm. The step is quite an important factor to be taken into account, because it determines if the descriptor is going to be extracted from every pixel or not. Figures 7 and 8, show this analysis in the Caltech and ImageNet dataset, respectively.

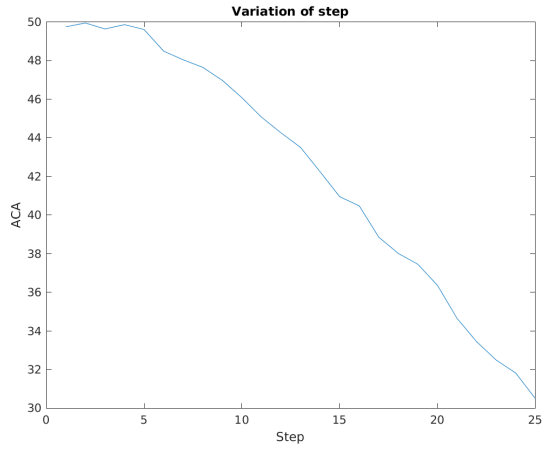

Figure 7. ACA when varying the step in Caltech.

As in the previous analysis, the response of both datasets is similar. These results show that the bigger the step, the less accuracy the algorithm will have in a range of 1 to 25 pixel step. This, is explained due to the fact that apparently, the PHOW descriptor in every pixel is necessary to obtain robust information that will be relevant to learn the vocabulary.

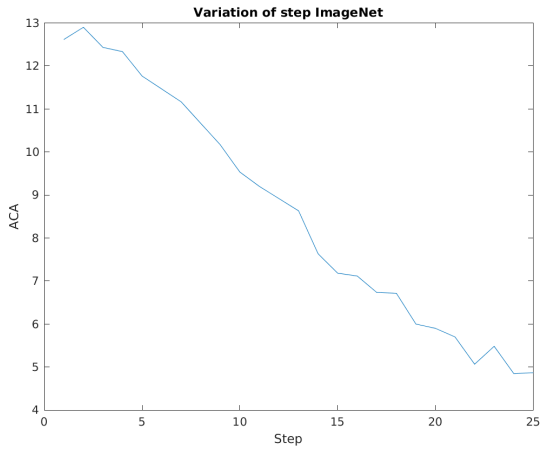

Figure 8. ACA when varying the step in ImageNet.

Due to the fact that each parameter was evaluated separately, the conclusion about the better ones can only be made this way. In the range analysed, the best parameters for the size of the bag of words is 275 for Caltech and 100 for Imagenet, altough this is not a relevant parameter, as explained before.

For the window size, the best values were 7 and 2, for Caltech and ImageNet, respectively. This difference can be explain because of the complexity in the ImageNet dataset in terms of background and occlusion, as can be analysed in figures 1 and 2.

For the step hyperparameter, the best results were 2 in both datasets. However, in figure 7, it can be seen that the ACA does not vary much during the first variation of the step for the Caltech dataset, contrary to the Imagenet, as seen in figure 8. The reason of this is the same one explained before about the complexity of the second set.

In general, the PHOW algorithm presents complication when exposed to problems where there are multiples objects in the image or when instances are occluded, and that is whats happens in the Imagenet data set. In figure 9, it is shown the confusion matrix of the best number of words parameter found in the Caltech dataset. It can be seen that the accuracy is 52.02, contrary to what is shown in figure 10. This shows the confusion matrix of the best window size hyperparameter found. Visually, the matrix does not show

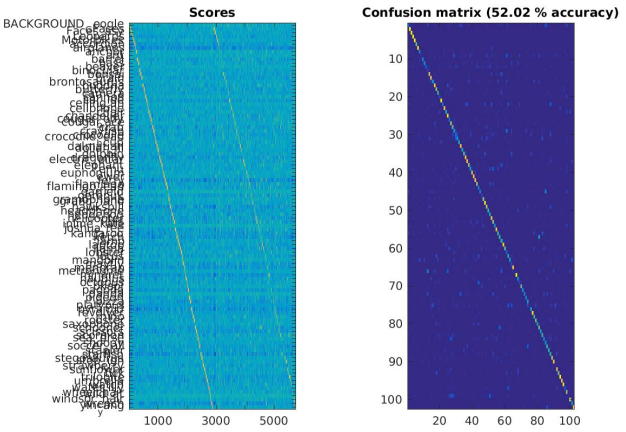

Figure 9. Best confusion matrix obtained in the Caltech dataset.

a good performance even in the trainning set, as the accuracy is only of 16.43. The reason for this has been already mentioned.

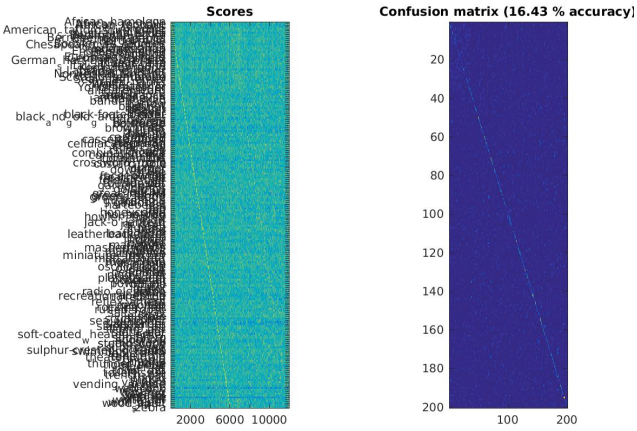

Figure 10. Best confusion matrix obtained in the ImageNet dataset.

Regarding the hardest class from each dataset, it was extracted from the analysis previously made, that the hardest one for Caltech is metronome. An example of this class can be seen in figure 11. This might be the hardest class because of the background of the image. The other images in this class also have different backgrounds with letters. This, might cause the descriptor to include information that misguides the classification task.

On the other hand, in figure 12, it is shown the hardest class in the ImageNet dataset: anemone. Visually, the

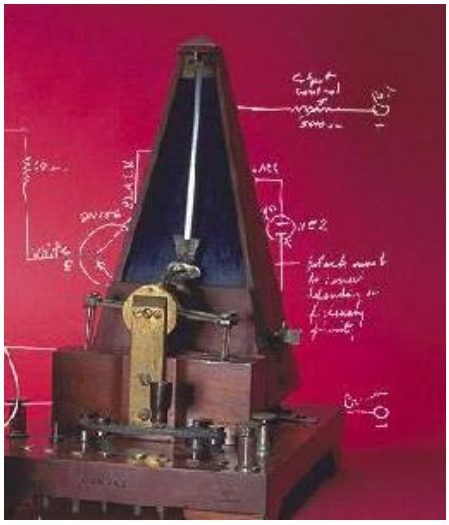

Figure 11. Hardest class, metronome, from CalTech.

reason for its difficulty can be identified. Each image contains different entities with various color, shape, and texture, which is not constant in every image. Thus, making the classification task harder.

Figure 12. Hardest class, anemone-fish, from ImageNet.

To obtain better results from the algorithm, it would be logical to include other features, as well as other descriptor to make the model more robust, under-minding the process time it would take .

References

[1] Vo T., Tran D., Ma W., Nguyen K. (2013) Improved HOG Descriptors in Image Classification with CP Decomposition. In: Lee M., Hirose A., Hou ZG., Kil R.M. (eds) Neural Information Processing. ICONIP 2013.

Lecture Notes in Computer Science, vol 8228. Springer, Berlin, Heidelberg 1

[2] A. Vedaldi, "VLFeat - Documentation ¿ C API", Vlfeat.org. [Online]. Available: http://www.vlfeat.org/api/sift.html. [Accessed: 24- Mar- 2019]. 1

[3] P. Rosado, E. Figuera, M. Planas and F. Reverter, "La visin artificial, un nuevo aliado para el anlisis de imgenes artsticas", 2015. [Online]. Available: https://revistas.ucm.es/index.php/ARIS/article/viewFile/48802/48284. [Accessed: 24- Mar- 2019]. 1

[4] 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 1

[5] "ImageNet", Image-net.org, 2016. [Online]. Available: http://www.image-net.org/. [Accessed: 24- Mar- 2019]. 1

[6] J. Hou, J. Kang and N. Qi, "On Vocabulary Size in Bag-of-Visual-Words Representation", Advances in Multimedia Information Processing - PCM 2010, pp. 414-424, 2010. Available: 10.1007/978-3-642-15702-8-38 [Accessed 24 March 2019]. 2