PHOW Classification in Caltech 101 and ImageNet

Luisa Fernanda Borbón R.
Universidad de los Andes
Departamento de Ingeniera Biomdica
lf.borbon@uniandes.edu.co

Abstract

Image Classification is an important computer vision problem in which a category is assign to an image based on its content. Traditionally, the Caltech 101 dataset has been used to solve this challenge but as its images present low clutter, other datasets like imageNet can be used to approach a more real life model. A model based on PHOW representation was implemented over this two datasets, using the VLFeat library developed by Vedalvi and Fulkerson. Varying its different hyperparameters and the number of train and test images, the maximum accuracy score value found for Caltech101 was 71.32% and for ImageNet200 was 21.35%. The importance of the hyperparameters, and main causes of failure in the algorithm are discussed, concluding that making a richer image representation with a hierarchical algorithm could help improve the results.

1. Introduction

Recognition problems in computer vision aim to obtain semantic information from local or global image features. This type of problem can be analyzed in different complexity levels which include: image classification, image annotation or tagging, object detection, object identification, image parsing and object parsing [1]. For instance, image classification is the first level of recognition problems in which a category is assign to an image based on its features. This problem is the basis to understand the information of the image as it gives an initial general approach to its content.

The most important datasets involved on image classification problems are Caltech 101 and ImageNet. The first was realead in 2004 by Fei-Fei, Fergus, and Perona; and include 40 to 800 images per each 101 categories (102 including background). On this dataset, images have little or no clutter and the objects tend to be centered and thus are relatively easy to recognize [3]. On the other hand, the imageNet dataset presents a wider range of images, with higher clutter and variability within each category. This dataset introduces higher difficulty to the problem of classification and because of this, can be seen as a problem closer to real life scenes, where objects don't appear clearly and centered but present in challenging ways such as different scales, positions, variable illumination and occlusion [2].

As mentioned before, in order to classify the image its features need to be extracted by using a rich representation, that will be able to enhance the relevant parts of the image or object of interest. A classical approach to this problem is the Pyramid Histogram of Visual Words (PHOW), which applies a dense scale invariant feature transform (SIFT) in order to construct a visual dictionary and classify an image according to the distribution of visual words that compose it. Despite its simplicity, this model has shown promising results, as presented by Lazebnik in 2006 [4].

In this sense, the present work intends to apply the PHOW representation approach to classify images from the datasets Caltech 101 and a subset of ImageNet200. This was done by using the PHOW algorithm written by A. Vedaldi and optimizing its parameters on the training set for each dataset [5].

2. Materials and Methods

In order to classify the images from Caltech 101 and the subset of ImageNet, as mentioned before the VLFeat open source library developed by Vedaldi and Fulkerson was implemented. On the following sections, the two datasets and the methodology followed are described in detail.

2.1. Datasets

The Caltech 101 dataset consists of 40 to 800 images for each 101 categories (Most categories have about 50 images) and the size of each image is roughly 300 x 200 pixels. Some of the categories included on this dataset are for example: airplane, camera, chair, cup, panda, strawberry and umbrella [3]. The images of this dataset were collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato; and set to public in 2004 in the Workshop on Generative-Model Based Vision. As mentioned before, most of the images tend to be centered and have no clutter; as is shown in figure 1 for the categories strawberry and

cup.

Figure 1. Sample of the caltech 101 Strawberry and cup categories.

On the other hand, the imageNet dataset is composed by millions of images, organized according to the WordNet hierarchy. As described on their website, each meaningful concept in WordNet, is used in imageNet as category called "synonym set" or "synset". There are more than 100000 synsets and the majority of them are nouns, with 1000 images on average for each category. Unlike the Caltech dataset, the images on ImageNet present high clutter and variability within each category, as shown in figure 2 for chihuahuas and folding chairs. Due to the large size of this dataset, only a subset of 200 train and test categories with 100 images was selected to apply the classification model.

Figure 2. Sample of the imageNet dataset for Chihuahua and folding chair categories.

2.2. PHOW Strategy

As mentioned before, to classify the images the VLFeat library was used to run the PHOW representation model. The image representation by Pyramid Histograms of Visual Words (PHOW) is an extension to the bag-of-words (BOW) model in which SIFT features are extracted from the images and treated as words.

The BOW model can be considered as the next step to textons, in which instead of finding local patterns from the results of passing the image trough 32 filters, the local patters will be representative sections of the image called visual words. As well as in the Textons representation, the image is then going to be understood according to the distribution of the patterns identified, in this case, how the visual words are presented on the image [1]. In this sense, in PHOW the image will be classified according to the frequency and distribution of the visual words presented, which may be significantly more effective in classification than using only textures to understand the objects.

Table 1. Default parameters used in Vedaldi and Fulkerson PHOW algorithm.

HyperParameter	Default Value
Train Images	15
Test Images	15
Number of Words	600
Spatial Partitions	[2 4]
SVM (C) Parameter	10

In order to extract the visual words from each image and construct the visual vocabulary, the dense Scale Invariant Feature Transform (SIFT) is executed over the training images. This is done by going through each pixel of the image, taking at window associated to the pixels surrounding it, dividing that patch into 16 cells and making the convolution between each patch and 8 orientation filters. As a result, a 128 dimension descriptor is obtained for each pixel, and using k-means to find the centroids of the vectors, the visual words are created.

In order to include spatial information PHOW uses the pyramidal representation by making this process in different levels. Finally, the SVM classifier is used to set the most similar category to the image of interest. As the SVM method is used for binary classification, many SVM models need to be trained and to help fasten this process, decision trees can be included.

2.3. Hyperparameters Selection

To estimate the best hyperparameters the approach followed was to vary each of them independently on the train set and calculate the accuracy score on the test images to evaluate the performance of the algorithm. The hyperparameters taken into account to vary and optimize were: the number of train and test images, number of words, the spatial partitioning and the c parameter (sometimes called penalty or confidence) for the SVM classificator. Other hyperparameters that could have been taken into account are more parameters associated to the classification methods like the number of decision trees and the kernel used in SVM.

3. Results

First, the PHOW strategy was tested on the Caltech101 dataset and using the default parameters, an aca of 68.10% was obtained. This initial parameters are shown in table 1

The best classification model was found by changing the hyperparameters and fixing the ones with that showed the best performance on the test set. First, the number of train and test images were varied between 20 and 100 as shown

in the figure 3. As the best performance was obtained with 20 train and test images, this number was fixed to evaluate the other hyperparameters.

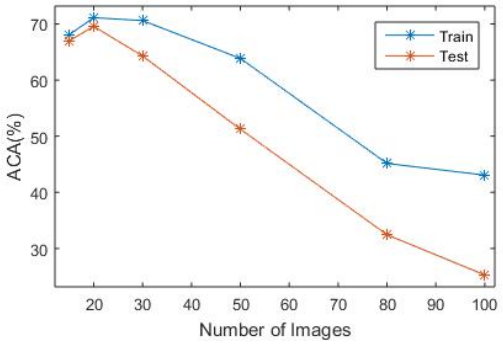

Figure 3. Accuracy variation when changing the number of training and testing images, the best aca was obtained with 20 images.

Next, the number of words was varied between 400 and 1200, obtaining the best performance as the number of words increased, as shown in figure 4. For the spatial partition values, the one which presented the best performance was [2 5], followed by [2 4 6] as shown in figure 5. Finally, the the value of the c parameter of the svm clasificator was varied between 0.1 and 15, finding as figure 6 shows, that the best result was obtained using the default parameter at c=10.

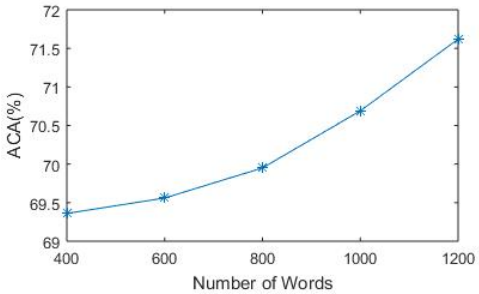

Figure 4. Accuracy variation when changing the number words taken into account.

In this sense, the best performance for classification in caltech101 was obtained with the hyperparameters shown in table 2. Using this parameter configuration, the best accuracy score obtained was 71.32%, and its corresponding confusion matrix is shown in figure 7.

After finding the optimal hyperparameters for the caltech101 dataset, the same procedure was done with the sample of ImageNet with 200 categories. First, the model was tested on the dataset using the default configuration as shown in table 1. Using this hyperparameters, the

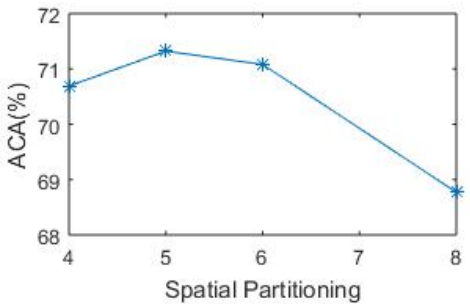

Figure 5. Accuracy variation when changing the value of the spatial partition taken into account.

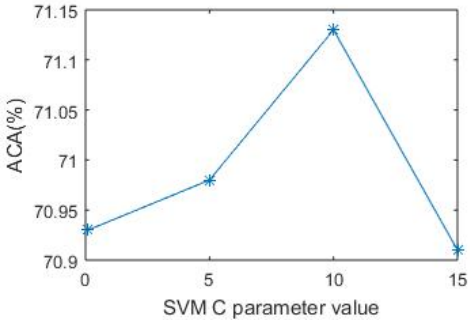

Figure 6. Accuracy variation when changing the value of the svm c parameter.

Table 2. Best hyperparameters found to classificate the Caltech101 dataset using the PHOW strategy.

Hyperparameter	**Best Value**
Train Images	20
Test Images	20
Number of Words	1200
Spatial Partitions	[2 5]
SVM (C) Parameter	10

maximum accuracy score obtained was 16%. By applying the best model configuration to this new dataset, the best score obtained was 17.2%, which is not significantly higher than the previous result. Finally, the hyperparameters were varied again in order to find a better result and using 50 Train and 20 Test images, 1000 Words, [2 5] spatial partitions and c=10 confidence in SVM Classificator. Figure 8 shows the confussion matrix associated to this model, which hyperparameter configuration was able to obtain an accuracy score of 21.35%.

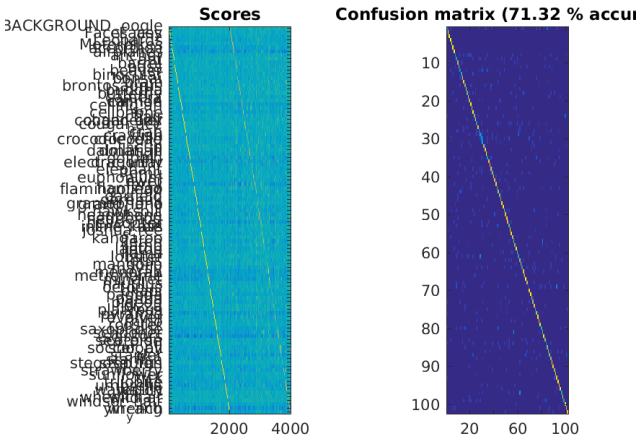

Figure 7. Confusion Matrix result for the best hyperparameter configuration in Caltech101 (20 Train and Test images, 1200 Words, [2 5] spatial partitions and c=10 confidence in SVM Classificator).

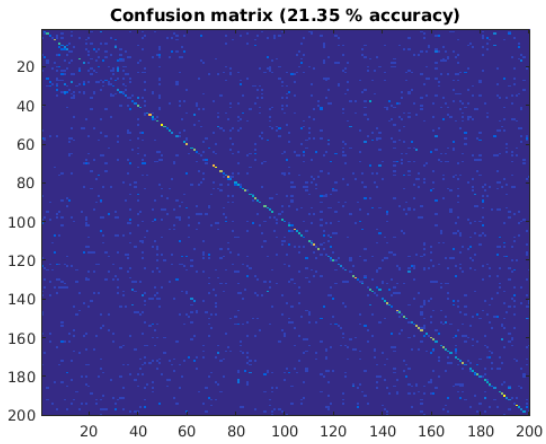

Figure 8. Confusion Matrix result for imageNet200 dataset using the best hyperparameters configuration.

4. Discussion

4.1. HyperParameters

Evaluating independently each hyperparameter reassured their importance and impact on the PHOW model implemented; as changing their value led to increases or decreases in the accuracy score of the algorithm. As the result show, the best hyperparameters vary according to the data analyzed, which is evident considering that the parameters define the model's adaptation to the data in order to generalize its characteristics.

The most significant hyperparameters were the number of train and test images and the number of words. Each dataset showed an optimal number of images to work with, which is associated to the complexity of the images wanted to be learn. Raising the number of train images in caltech beyond 20, lead to overfitting or the data and thus to a decrease in the test accuracy score. On the other hand, even thought increasing the number of words shown an increase in the accuracy score, this amount should be controlled and tested as it may lead to over-fitting of the model. Specifically, increasing the number of words shown to increase the accuracy of the algorithm but it also increased the training time, and because of this, no measurements beyond 1200 words were made.

4.2. Categories

The PHOW representation enhances different characteristics of an image, and due to the high variability of each category presented, the fact that some images were classified easily o not classified at all was expected. A sample of some of the best and worst classified categories is shown in figure 9, and possible causes of this results will be described next.

For instance, the easier or best classified categories were: website (80%), house_finch (75%), alligator lizard (70%), coral fungus (65%) and dowitcher (65%). The high accuracy of the model in the website category can be explain by the fact that it doesn't present high variability in terms of clutter and illumination. Compared to the other categories, website images have a really defined structure with a lot of horizontal and vertical borders which can be representative when extracting the dense SIFT features and result in a better classification. In addition to this, two of the higher rated categories were birds (house finch and dowitcher), which can be associated with the low occlusion of this images as this photos focus on the bird and the characteristic shape and texture they have. Even though they are birds, this two breeds have an special shape like a larger peak or tail, which make them different from species and may contribute to the fact that they are easy to identify. In regard to the alligator lizard and coral fungus categories, their good performance can be explained by the fact that they also have unique differentiate shapes and textures, the coral fungus for example has a vertical texture organization that is not presented of similar to other objects of categories.

On the other hand, there were 26 categories that didn't have any correct classification. This worst categories (Aca 0%) on the imageNet dataset included american staffordshire terrier, blenheim spaniel, border collie, box turtle, labrador terrier, comodo dragon and indri. Most of this categories included animals, and recurrently this categories were associated with dog breeds. The low performance in this category may be explained by the fact that in general dogs have a similar shape and thus it can be difficult to extract significant features that represent a single breed. Also, when looking at the images it was evident that the level of

variability within the classes is really high, as shown for the Chihuahua category in figure 2. As well as the box turtle, comodo dragon and indri categories, most images presented had high levels of occlusion of the objects of interest with outdoor nature and plants, indoor furniture, people or other animals; with may have contributed to low performance obtained on this categories.

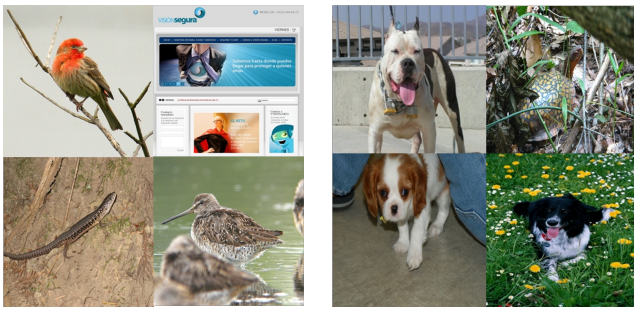

Figure 9. Sample of the imageNet best(Left) and worst(Right) categories classified by the best model.

4.3. Challenges and Improvements

The main challenges identified were mentioned before, for instance the great variation within each category and shape similarities between objects in different categories may lead to errors in the PHOW model. In order to improve the results, other hyperparameters of interest can be varied, to start, changing the SVM kernel might be a good idea in order to make a more accurate representation of the data. Other important aspect found was the great proportion of occlusion, illumination, and cluttering challenges that were encountered within the imageNet dataset categories. This challenge may be addressed by doing a rigorous image prepossessing or subsamplig the images of the category by automatically selecting the ones that present less classification troubles.

To address the problem found when classificating dog breeds, some kind of hierarchical algorithm could be done in order to take the whole group of dog breeds and start dividing them according to smaller local characteristics to improve the PHOW algorithm.

5. Conclusions

The PHOW algorithm presents itself as an effective method to classify images into a large number of categories, when there it is accompanied by a good enough representation method. The overall performance of the classification algorithm was significantly lower on the imageNet dataset in comparison to caltech101 images as the first one has higher complexity level images that present challenges like occlusion, clutter and high variation within categories.

Also, the hyperparameters need to be adjusted specifically for each group of data and taking into account the processing time, as increasing the value of an specific parameter may increase the accuracy score but also the computing time and required computational power.

References

[1] P. Arbelez. Lecture 12: Recognition 01, computer vision ibio4680, 2019. Universidad de los Andes.

[2] ImageNet. About imagenet, 2016. [Online] http://image-net.org/about-overview.

[3] R. F. L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, 2004. [Online] http://www.vision.caltech.edu/feifeili/Fei-Fei$_GMBV04.pdf$.

[4] J. P. Svetlana Lazebnik, Cordelia Schmid. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, 2006. [Online] https://hal.inria.fr/inria-00548585/document.

[5] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.