PHOW

Federico Ulloa
Universidad De Los Andes
Cra. 1 18a-12, Bogotá, Colombia
f.ulloa10@uniandes.edu.co

Esteban Vargas
Universidad De Los Andes
Cra. 1 18a-12, Bogotá, Colombia
e.vargas11@uniandes.edu.co

1.. Introduction

Classification is one of the main areas in computer vision. It consists in assigning images to predetermined categories. Normally, the average classification accuracy (ACA) is used to evaluate classification algorithms. This metric is obtained by constructing the confusion matrix, normalizing it, and calculating the average of the diagonal.

The Caltech 101 dataset in 2003 in the California Institute of Technology by Fei-Fei Li and his team. The dataset consists of 9.146 images divided in 101 object categories (hence, the name) and one background clutter category. Each category is composed of around 40 to 800 images, and each image is about 300x200 pixels. Some oriented objects, such as vehicles, are often mirrored or rotated to provide more variants. [1]

The ImageNet dataset is a large database that currently contains over 14 million URL's of hand annotated images. It has over 20.000 categories, and each category has an average of 500 images. All these images are not owned by ImageNet and are available to everyone. Due to the size and origin of all these images, they vary in size and the categories are unbalanced. [2] For this laboratory, only a subset of 200 classes with 100 images per class is used.

The Pyramid of Histograms of Visual Words (PHOW) is a way to make image description. It is an extension of the Bag-Of-Words (BOW) model in which the SIFT features extracted are treated as words. PHOW is the combination of the combination of the space pyramid model and the bag of features. [3]

In this method, SIFT is first applied. SIFT (Scale Invariant Feature Transform) is a method in which important features of images are extracted to describe the image. Then, a dictionary is generated from these features using k-means clustering. Finally, a space pyramid is performed to compensate the fact that SIFT ignores spatial and structural information. [4]

From this, we can say that the difference between SIFT and PHOW is that SIFT is simply the extraction of features of the images, while PHOW uses SIFT first, then it uses k-means to create a dictionary and then it uses spatial pyramids to include spatial and structural information. Also, as PHOW is based in SIFT means that it is also scale invariant, as the descriptors extracted are relative to the key point detection scales. [5]

2.. Methodology

The provided script that implements the PHOW method was used to evaluate both the Caltech 101 dataset and a subset of the ImageNet dataset.

First, for the ImageNet dataset, experiments were run varying the number of words in the dictionary, the size of the train-set, the C of the SVM and the spatial partitioning. In order to find the best combination of parameters, an exhaustive search was made, is important to note that the code takes a lot of time to run, so making a parameter mesh and try all the possible combinations is a really time consuming task, therefore, all the parameters minus one were fixed in order to analyze the influence of each parameter in the ACA. The same methodology was implemented to find the best hyperparameters in the Caltech-101 dataset.

Consequently, not all the hyperparameters were iterated. Only the ones mentioned previously were varied, as those were the ones that we found had the most influence in the results.

3.. Results

The results are divided into ImageNet dataset results and Clatech-101 dataset results

3.1.. ImageNet-2000 Dataset

First, the number of words in the dictionary were modified, leaving the other hyper-parameters fixed, C=10, training set size=15, spatial partitioning X/Y = [2 4].

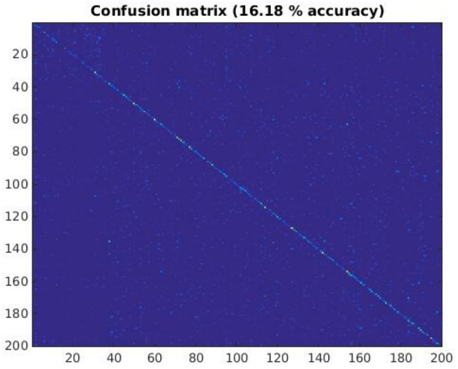

Figura 1: Number of words in dictionary=600

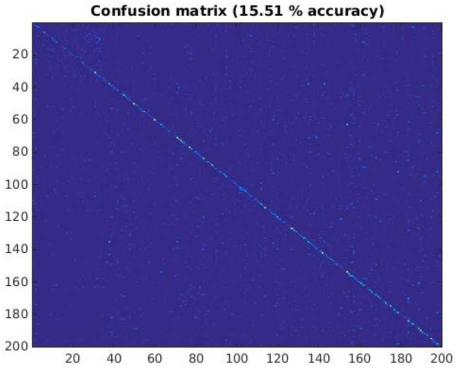

Figura 2: Number of words in dictionary=400

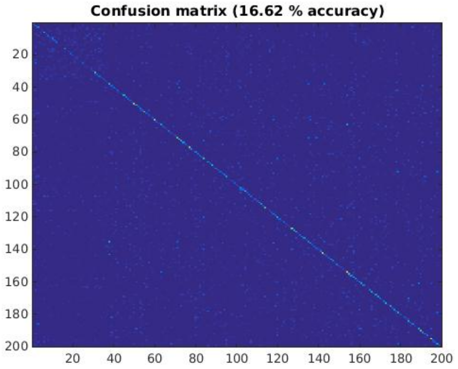

Figura 3: Number of words in dictionary=600

As expected, when the number of words is increased the representation is more descriptive (Because there are more features) so the performance of the classifier (ACA) is increased by 1 percent. After an 800 number of words, the performance of the algorithm didn't increase but the computational time increased, so 800 words were selected.

Then, hyperparameter C (Regularization term of the SVM) was modified leaving the other hyper-parameters fixed, number of words =600, training set size=30, spatial partitioning X/Y = [2 4].

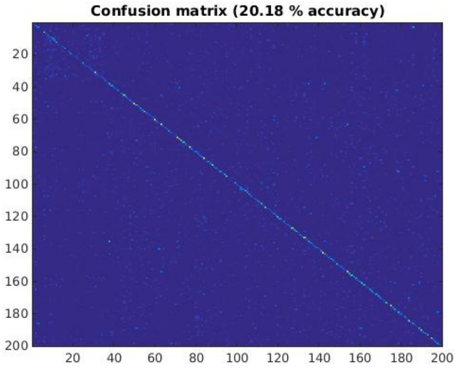

Figura 4: C=10

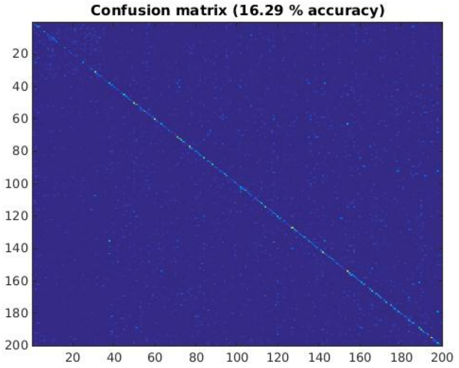

Figura 5: C=20

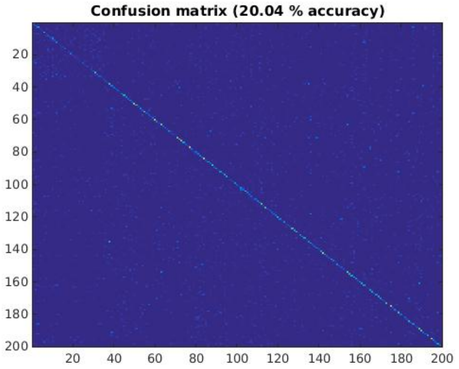

Figura 6: C=5

First, the overall performance of the classifier increased when the train dataset size increased from 15 to 30, which was expected because when the train set is increased the algorithm is able to learn different representations of the same class.

On the other hand, C=10 shown to be the best regularization term, taking into account that with C=5 the algorithm probably have a high variance, and with C=20 high bias.

One experiment was done with a different spatial partitioning [1 2] which means that in the histogram pyramid, the first histogram correspond to the full BOW histogram and the second one is the concatenation of the two half image histogram. C=10, number of words=600 and train size=30 were used.

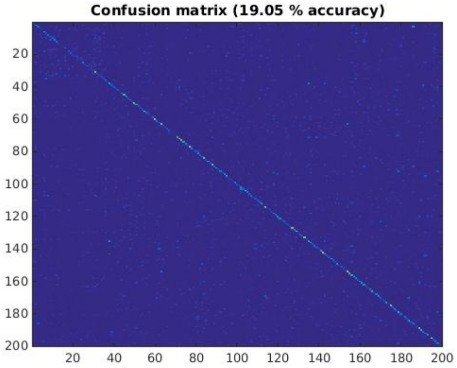

Figura 7: Spatial Partitioning = [1 2]

The performance of the algorithm decreased by 1 percent with respect to 4 which have the same hyper-parameters expect for the spatial partitioning.

Finally, C=10, train size=50, number of words=800 and spatial partitioning [2,4] were selected as the best hyper-parameters.

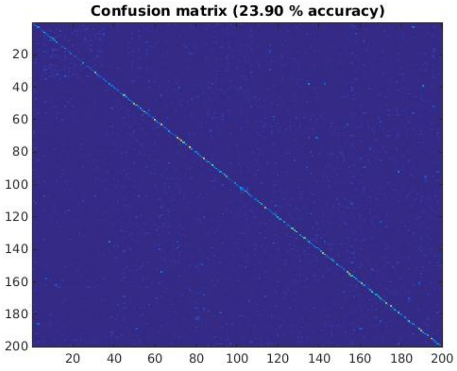

Figura 8: Best hyper-parameters result

The most difficult classes, were

1. Weasel
2. Labrador-retriever
3. Great-dane

On the other hand, the easiest classes were

1. Website
2. Brass

Is is notable that, the hardest classes for he classifier were classes with animals, probably because of the different background. Also, the easiest classes for the classifier were classes where the background is more or less the same as in the websites.

3.2.. Caltech-101 Dataset

Making the same analysis than in the imageNet-2000 dataset, the best hyper-parameters found were C=10, number of words = 800, training set size = 30

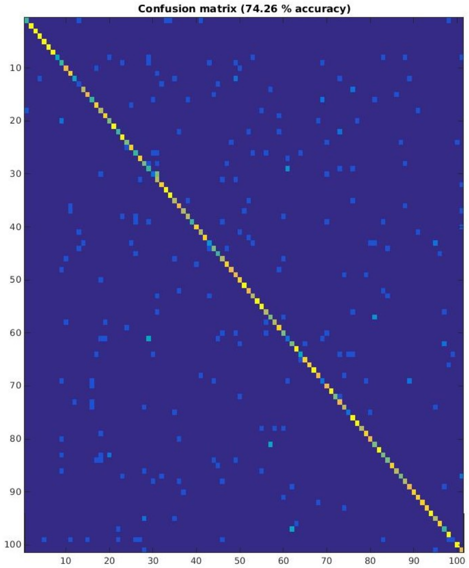

Figura 9: Best hyper-parameters result

An 74.28 percent of ACA were the performance of the classifier, comparing this result with the results in the imageNet dataset 23.9 percent, is notable that the complexity of caltech-101 dataset is less than the imageNet dataset. This is because the caltech data base have objects in similar spatial conditions while the imageNet dataset is more variate with really different environment conditions for the same object.

4.. Conclusions

After analyzing the influence of different hyperparameters, we found that, first, as the number of words increases, the representation is more descriptive. This works until a point in which performance level no longer increases, and only computation time is increased, meaning there is an optimal number of words, which was 800 in this case for ImageNet dataset.

Regarding the train dataset size, performance also increased as this increased, as the classifier can learn more representation of the same classes.

The C in the SVM showed that it also has an optimal point, as with a low C, the classifier has a high variance and with a high C, it has high bias. In ImageNet dataset, the optimal C was found to be 10 for a better performance.

Also, we found that spatial partitioning alters the performance of the classifier. Changing from [2 4] to [1 2] decreased the ACA obtained.

Furthermore, performance is different among different classes, as within each class there is a different variance. This meant that there were classes easier to classify than others.

Lastly, a big difference in the ACA was obtained in the 2 datasets evaluated, with a way bigger one in the Caltech 101 dataset. This happens because Caltech 101 dataset has a lot less variance in some of their features, while this is quite more complex in ImageNet dataset.

To improve the method, convolutional neural networks could be added. This has been found to be a very effective method for image classification and it is broadly used, and it is an area that is still in big development stage. [6]

Referencias

[1] Çaltech101", Vision.caltech.edu. [Online]. Available: http://www.vision.caltech.edu/ImageDatasets/Caltech101/. [Accessed: 20- Mar- 2018].

[2] [Online]. Available: http://www.image-net.org/about-overview. [Accessed: 20- Mar- 2018].

[3] Shereen A., HowydaYoussryAbd E., Aliaa A. Image Multi-Classification using PHOW Features. [Online]. Available: http://webcache.googleusercontent.com/search?q=cache:https://pdfs.semanticscholar.org/9e2e/972d186aa543facf5ec50c171642a690f08a.

[4] Wang J., Yan H., Li J., Xia P. PHOW Based Feature Detection For Head Pose Estimation.

[5] [Online]. Available: https://www.quora.com/Why-are-SIFT-descriptors-scale-invariant. [Accessed: 20- Mar- 2018].

[6] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review", Neural Computation, vol. 29, no. 9, pp. 2352-2449, 2017.