Laboratory 8: PHOW Classification

Javier Coronel
Universidad de los Andes
Biomedical Engineering
jd.coronel30@uniandes.edu.co

Luis Carlos Rivera
Universidad de los Andes
Biomedical Engineering
lc.rivera10@uniandes.edu.co

Abstract

PHOW classification also know as Pyramid Histogram Of Visual Words is a methodology of deep learning techniques based on the bag of words model. In this models, a feature is treated like a word, and the main idea is to classify an image based on the frequency of the presented features which fits the best to the features of a certain category. In the following report will be analyzed the performance of this methodology over the ImageNet Dataset. The results was as expected, as the number of categories increases, the computational model required to create the SVM multi-class model was bigger and the average classification accuracy decreased very fast.

1. Introduction

Pyramid Histogram of Visual Words (PHOW) is a classification model witch use features presented in an image as words that may represent the image content. This methodology was first implemented in text understanding, the features will be represented in a sparse vector associated with the frequency of each of the features. The main issue with this methodology is the lost of spatial information associated with the original image, in general the model know that there is a feature somewhere in the image and the number of times witch that feature appears over the image [1].

To analyze the behavior of this methodology over a big and diverse dataset the ImageNet dataset will be used . This datasset is composed of 996 different categories, for each of the categories we count with training data and test data, the number of images is equal for each category and all categories have 100 images in the training set and 100 images in test set. All the images are in RGB and for spatial normalization over the dataset all the images had been re-sized by default to 256X256 pixels.

The main idea of the present report is to analyze the performance of this methodology over a big dataset of images that have many different categories. To identify how the algorithm behaves, as the number of categories increases, the computational time required and the average accuracy from each model in training and testing set will be inspected..

2. Methodology

For this laboratory we used the vl-library witch is open source,this library has tools specialized for image understanding, local features and matching [2]. They tested several of their algorithms on different datsets, but for this specific case we implemented their test of PHOW in caltech-101 in a different dataset with more categories like is ImageNet. The changes made over the algorithm to work in this dataset and analyze the performance were the following:

- The number of categories, for this specific case will be the same number of categories of training and testing. We implemented the code to iterate the number of categories increasing each step by 10 categories until 240 and 120 categories are reached for training and test respectively.
- The number of clusters in k-means classification of the features, this number increased in order to get a better classification of the features. This because of the big amount of categories, and some of this may not differ so easily if there were a reduced number of descriptive features.
- Images included in training, the number of images of training taken into account to make the learning model. It was changed to identify its effects on the accuracy classification performance. The variations were identified over a classifier of 50 classes and the number of images selected were increasing in 5 each time.
- Number of spatial partitions, the number of spatial partitions of the image to obtain the histograms and the feature representation, changed to identify its effects on the accuracy classification performance. This effects were identified by setting the spatial partitions in 2 and increasing until 10 in the same previous classifier of 50 classes.

- In the phow-caltech-101code, they find and used all the images in .jpg format. In ImageNet the images are in .JPEG format, so it was necessary to change the file format searched by the code.

3. Results

3.1. Training

In figure 1 is presented the average classification accuracy over ImageNet dataset as the the number of categories increases. In figure 2 we present the computational time required to create the SVM multi-class model as the number of categories increases.

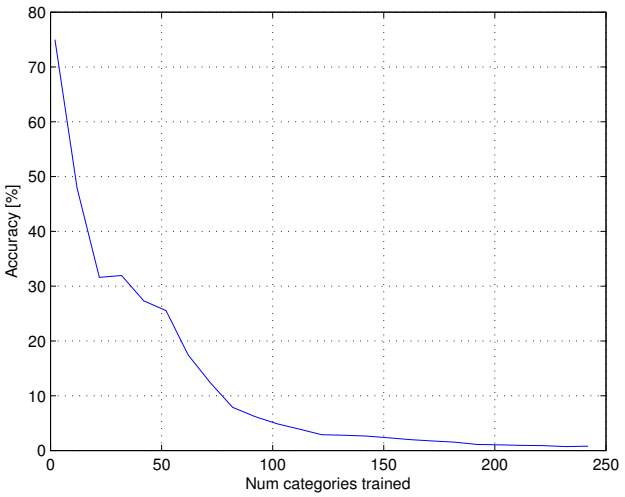

Figure 1. Accuracy variation depending on number of categories included.

In figure 3 is represented the accuracy variation depending on the spatial partitions of the image on a 50 classes classifier, the spatial partitions in X and Y are the same.

In figure 4 is represented the accuracy variation depending on the number of images included for training.

3.2. Testing

The performance over the testing set was evaluated using the same number of categories trained in the previous section. We also evaluated the average classification accuracy and the time required to predict the labels as the number of categories increases. The Figure 5 represents the behavior of time as the number of categories increases in the model. It is important to mention that we tested the model over all the categories presented in the model. The Figure 6 presents the accuracy of the method as the number of categories increases.

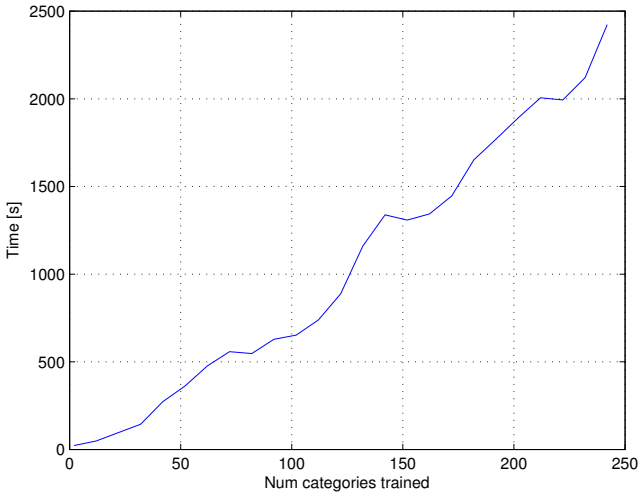

Figure 2. Time variation depending on number of categories included.

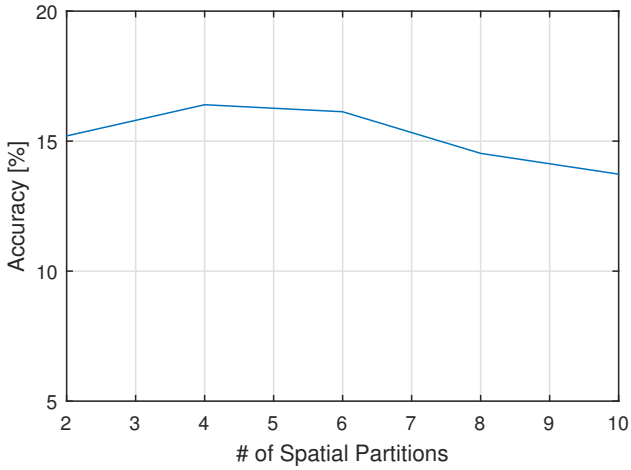

Figure 3. Accuracy variation depending on spatial partitions.

4. Discussion

In Figure 2 it is noticeable a linear behavior of the required time to made the classification, it means that, the more classes are included for classification, the greater is the required time to classify.

In a similar way, Figure 1 shows that for a larger number of classes, the average classification decreases exponentially.

Regarding to the number of training images included for the model, it is noticeable that there is an improvement in the classification accuracy. And this is reasonable because a wider description of the classes is reached when there is more images to learn from. Thus, Figure 4 shows an improvement of almost 7% in the training set, despite of this it is important to have in mind that include a larger number

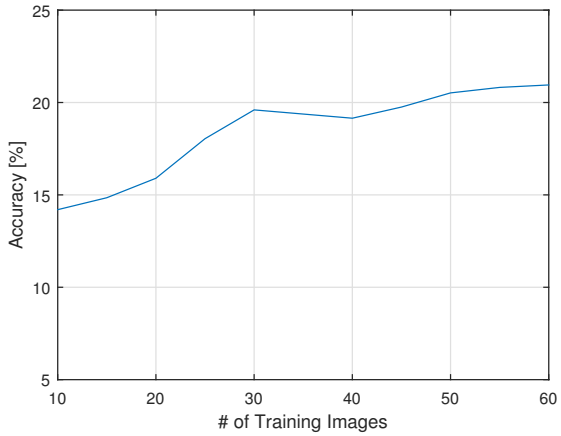

Figure 4. Accuracy variation depending on number of images included on training.

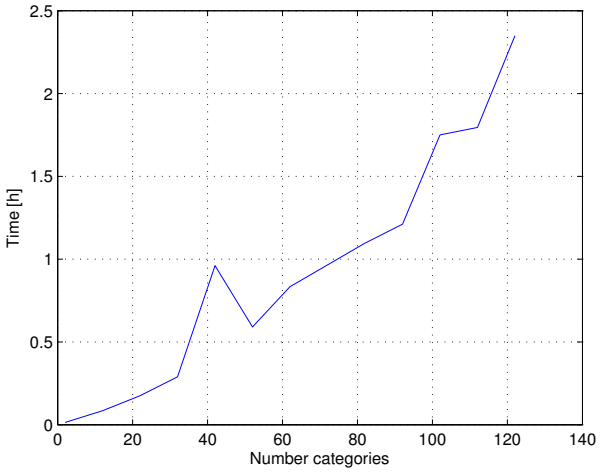

Figure 5. Time variation depending on number of images included on test.

of images could result in overfitting for the model.

On the other hand, for the spatial partitioning of the images the Figure 3 shows an improvement of almost 1.5% but the behavior of the change seem to don't have a significant effect over the classification accuracy.

For the test, in figures 5 and 6 is shown a similar behavior as the training, where a exponential decay is in the accuracy and a linear behavior in the required time as the number of classes increases.

In general terms, with this method it is evident a lack of accuracy for the classification. This might show the problems of migration of methods in different datasets. Making really hard to identify at which point the model is overfitting or underfitting for a given amount of images and categories.

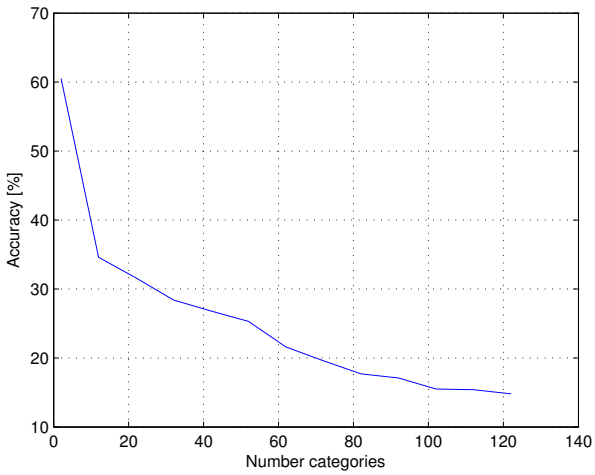

Figure 6. Accuracy variation depending on number of images included on test.

5. Future Work

A future improvement over this dataset could be not to train in order the categories but to identify witch of the categories have more confusion and train binary SVM with that categories,kind of bootstrapping but with the categories with the highest rate of confusion.
Another way to improve this results based on the computational time required by the training could be to determinate if the number of pictures of category are more than the needed, enough or less, because as the number of categories increased the computational time required increased also exponential as saw in the results section figure 5. Another way to try to improve this method is using binary SVM which will create a model for each pair (or less categories per model) of categories and in the end will set the label to the predicted label with highest score.

References

[1] S.-M. Khaligh-Razavi. What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. *arXiv preprint arXiv:1407.2776*, 2014.

[2] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1469–1472. ACM, 2010.