Image classification with PHOW strategy on Caltech-101 and ImageNet datasets

Catalina Gómez
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
c.gomez10@uniandes.edu.co

Diana Herrera
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
ds.herrera10@uniandes.edu.co

Abstract

Classification is one of the main problems in the Computer Vision fields, the goal is to assign a label to an image according to the content it has. Nevertheless, the amount of classes to classify objects in the world is as large as the language categories that can be found on a dictionary and it is almost impossible to use them all. Therefore, the number of classes used, the amount of images used to train a classifier and the amount or images from each class to test the algorithm are only some of the most important parameters in any classification algorithm. To tackle this problem, Andrea Vivaldi developed the PHOW algorithm, evaluated in this opportunity on the Caltech 101 and ImageNet datasets varying the most important hyperparameters. Performance was measured in terms of ACA and time and results obtained are considerably better for Caltech.

1. Introduction

The problem of recognition studied in the Computer Vision field pretends to describe the elements within an image. One of the main tasks is to assign a label to each image that corresponds to a specific category or class, or in other words, recognize categories of visual objects. However, there could be within-class appearance variations that make this a challenging problem to study, in addition with other aspects such as background clutter, scale, occlusion, changes of illumination, and the projection of a 3D world into a 2D image.

The general idea is that based on the visual appearance of the objects, the algorithm can recognize the category of the image, and these categories are derived from our language. In this way, it is possible to establish an organized hierarchy from language and visual categories, and study recognition at different levels.

In order to study this problem, a dataset called Caltech-101 was developed in 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The dataset consists of pictures belonging to 101 categories (102 including background), with about 40 to 800 images per category (most categories have about 50 images). The resolution of each image is 300 by 200 pixels. In addition, they provide annotations of the objects found in these pictures to evaluate the performance of algorithms of recognition [1].

Latter, in 2011 a new dataset was created for a large scale visual recognition challenge. With large scale it means that it contains millions of images and thousands of categories. This image dataset was organized according to WordNet hierarchy (including only the nouns), which is a large lexical database of English. Each *synset*, defined as a meaningful concept in WordNet described by multiple words or word phrases, is illustrated by almost 1000 images, and there are about 100,000 synsets. The images of each concept are human-annotated [2]. The leaves of the hierarchy correspond to categories with finer distinctions among classes. The evaluation metric can be performed over different levels of the hierarchy, according to the distinctions that want to be made.

A classic strategy for object recognition is known as Pyramid histograms of visual words (PHOW), which is a feature descriptor for object categorization. This algorithm is based on the scale invariant feature transform (SIFT), which works for category recognition. For a better discriminative power, it is better to use higher dimensional features (dense features) since they have shown to work better for scene classification [3]. That is why PHOW strategy computed dense SIFT on the training images. The main idea is to compute histograms of gradient orientations (8 reference angles) for each pixel inside a sub-patch, that is why it corresponds to local shape descriptors. Then, a visual dictionary is constructed with k-means clustering algorithm on SIFT space, where each cluster is a visual word. A pyramid of visual word histograms is computed for each image, and finally a SVM is trained for each category with intersection kernel. For the testing data, the pyramid of visual word histograms is computed and this is the input for all the SVMs. The predicted category with highest confidence is selected.

2. Materials and Methods

The VLFEAT open source library was used for the classification of images from different categories within the datasets of Caltech-101 and ImageNet. PHOW strategy was implemented with a function developed by Andrea Vedaldi, where it uses VLFEAT to construct an image classifier on Caltech-101 data. Later, this function was adapted to train a model for ImageNet dataset and then evaluate its performance on the test set. The most relevant parameters that were changed for the PHOW strategy in Caltech-101 dataset included the number of instances used in the train and test sets, the number of classes selected to train the model, the number of words used for PHOW, the C penalization parameter for the SVM classifier and the spatial X and Y to sub sample the image on the pyramid and determine the number of levels.

For Caltech dataset, the following scenarios were evaluated in order to compare and choose the most appropriate parameters for the dataset:

Table 1. Scenarios evaluated in Caltech 101

Default Scenario	
Num Train Images Num Test Images Num Classes Num Words Num Spatial X and Y SVM C	15 15 102 600 [2 4] 10
Scenarios with different Num Classes	
Num Classes	20 and 50
Scenarios with different Num Train Images	
Num Train Images	8 and 30
Scenarios with different Num Test Images	
Num Test Images	8 and 30
Scenarios with different Num of Words	
Num Words	300 and 900
Scenarios with different SVM C	
SVM C	5, 20 and 30
Scenarios with different Spatial X and Y	
Num Spatial X and Y	[1 2] and [4 8]

For ImageNet dataset, only three scenarios were evaluated by varying the number of train images used and then, classifying test dataset with the trained model. Three models were trained, one of them using 30 images for each class, a second one using 50 for each, and finally with the whole train set. Using these models, evaluation was made with 100 images per class (whole dataset) and with 50 images per class. In all scenarios, all the default parameters used were the same as in Caltech 101, except for number of classes, which is 200 for ImageNet dataset. The ImageNet training models were run on a machine with high computational power.

3. Results

Results for ACA and time in minutes are shown for all the scenarios evaluated in Caltech 101.

Table 2. Results from scenarios evaluated in Caltech 101

Default Scenario			
Tiempo 43 min		ACA 68.24%	
Scenarios with different Num Classes			
20		50	
Tiempo 9 min	ACA 80.67%	Tiempo 17 min	ACA 70.67%
Scenarios with different Num Train Images			
8		30	
Tiempo	ACA	Tiempo	ACA
24 min	58.63%	57 min	70.72%
Scenarios with different Num Test Images			
8		30	
Tiempo 31 min	ACA 67.77%	Tiempo 46 min	ACA 63.95%
Scenarios with different Num of Words			
300		900	
Tiempo 30 min	ACA 67.06%	Tiempo 37 min	ACA 67.25%
Scenarios with different SVM C			
5		30	
Tiempo 40 min	ACA 68.17%	Tiempo 39 min	ACA 68.10%
Scenarios with different Spatial X and Y			
[1 2]		[4 8]	
Tiempo 39 min	ACA 63.92%	Tiempo 48 min	ACA 66.14%

Similarly, results for ACA and time in minutes are shown for all the scenarios evaluated in ImageNet, which include the models with different number of train instances and varying the number of testing images.

Table 3. Results from scenarios evaluated in ImageNet

Varying Number of Train Images per Class					
30		50		100	
Tiempo	ACA	Tiempo	ACA	Tiempo	ACA
6 min	15%	10 min	18.8%	23 min	22.47%
Varying Number of Test Images per Class					
30 Image Model Trained		50 Image Model Trained		100 Image Model Trained	
50 Test Img	100 Test Img	50 Test Img	100 Test Img	50 Test Img	100 Test Img
17.1%	16.5%	18.9%	18.6%	**23.4%**	23.1%

The best result was obtained when training the model with the whole dataset, and when evaluating with only half of the test images. This was an average classification accuracy of 23.4% for 200 classes, which is much better than chance.

4. Discussion

It is noticeable that testing different parameters for each dataset can generate a difference in the overall performance of the algorithm, measured in ACA and time. Nevertheless, some parameters do not generate an important difference these measures, such as the number of words and the SVM C parameters in Caltech dataset. This might be because the number of words used are enough for the dataset and increasing the value further might lead to overfitting, a result that can be seen when comparing the default scenario (600 words) with the ACA obtained when using 900 words, which is smaller than the previous one. The same happens if SVM C parameter is increased too much, because then the error on the margin allowed is highly penalized, and this is noticeable with the values tested: when using the highest (30), ACA decreased compared to the default scenario, which used SVM C=10, but ACA with the default parameters is larger than the one obtained when SVM C was 5. Care must be taken when choosing this parameter of the SVM since an increased C could lead to overfitting.

On the other hand, some hyperparameters do generate an important increase or reduction in ACA in Caltech dataset. The number of classes is the one with the biggest effect on time taken, generating an almost two-fold increase when the value was duplicated. This result is completely expected considering that each class adds a significant amount of images to be represented, taken into account to train the classifier and then, to be tested. It is relevant to note that when the number of classes is increased the ACA decreases, which is perfectly logical because using more categories increases the probability of confusing the images belonging to each. That is why using less classes gave us the best ACA in the whole table with the shortest time. The number of train images had also an important effect in both time and ACA; increasing the number of instances increased the time cost because more images had to be processed and included to train the model, but they also increased the ACA, which is logical considering that the model had seen more images belonging to each class and hence, more features could be learned to improve the classification results. The opposite happened when changing the number of test images used to evaluate the model related to ACA results because when more images have to be classified in the test stage, there is more probability to mismatch their tags and therefore ACA decreases. In addition, time increases because anyway, more images have to be represented. Lastly, varying spatial X and Y parameter, corresponding to the sub sampling done, has a little effect on both time and ACA. When a bigger sub sampling is done time increases slightly, which is expected because more calculations must be done in order to compute the different levels of the pyramid representation, and ACA increases slightly too, a result that might be caused because sub sampling allows to recover local information concerning a window, not only pixels themselves.

Considering all the information, it is paramount to note that the best set of hyperparameters for Caltech 101 dataset must be chosen depending on the computational power available, but specially making a balance between ACA and time. Small increases in ACA might be obtained with higher values from parameters such as number of train images or spatial X and Y, but time cost could be too expensive for such an small improvement in results. Additionally, to improve ACA results significantly cheating can be done through reducing the number of classes of the problem or the number of images on test stage, but that would not be ideal for a serious algorithm that wants to be compared with the state of the art on classification tasks. Therefore, we conclude that for the scenarios tested and to truly evaluate the power of PHOW algorithm we would use the whole 102 classes, 30 train images and 15 test images, 600 words, SVM C=10 and Spatial X and Y = [4 8].

For Imagenet the results are similar, because increasing the number of images used in training, increased the time needed as well as the ACA, but not as dramatically as in Caltech 101. It is also evident that ACA obtained with default parameters is significantly smaller in this dataset when compared to Caltech. This might be due to the fact that ImageNet has twice the amount of categories than Caltech, which lead to an increment in the probability of mismatching the label of an image, and also because images are directly taken from the Internet, making them more complicated, varied and in general, less standardized when compared to the ones in Caltech 101. Nevertheless, results in ImageNet for test and train stage when using the models trained with 50 and 100 images did not presented a marked difference (only about 5% better when training with 100 images) and were even better for the testing set than for the training set, an unusual result in these supervised algorithms. This demonstrates that PHOW algorithm is generalizable and very useful to classify images that it has never seen by training a classifier with a big amount of images of each category. It may be important to highlight the effect of computational power on the required time to complete train and test tasks on the datasets, because as ImageNet always needs considerably more time than Caltech, the time in the second one with default parameters is four-times the one used for ImageNet, only because the machine used for the last is massively better than the one used for the first.

By comparing the ACA results on the ImageNet dataset, it was again confirmed the fact that increasing the number of images per category to train the model increased the its performance on the testing set, hence reducing the error, since the model can learn from a more general case. However, a significant increase in the number of the training instances could lead to the case where there is so much intra-class variation that the model can not learn accurately to predict the label. This was not the case of the datasets that were used to test PHOW strategy, so we did not faced the problem of increasing the number of training instances. In order to reduce more the testing error with many training instances, it would be a good strategy to increase the complexity of the model by optimizing the parameters, but having in mind that too complicated models can result in overfitting. As in Caltech-101, increasing the number of test images used to validate the models (with different training instances) resulted in a decreased ACA because the probability of mismatches and errors is increased.

Furthermore, when giving a closed sight to the results obtained for each class con Caltech 101, it is noticeable that classes such as faces, leopards, motorbike, airplane, car side or cellphone have a high correspondence value, so they can be considered easy because they tend to stand out from their background and are not occlude. On the other hand, classes such as Google Background, anchor, cannon or beavers have a very low score and this might be because their colors mix up with their backgrounds, the class is too general (Background), or their shape might be mistaken with objects from other classes (cannon).

The trained model that yield the best result on the test set of ImageNet was further analyzed in order to distinguish the classes that were classified with a higher accuracy. The top 5 five included the classes of website (76%), carbonara (74%), dowitcher (70%), theater curtain (70%) and coral fungus (68%). A possible explanation is that, for instance, in the case of carbonara class, the images were almost uniform (as in website category) since they were mainly a dish filled with pasta. The dowitcher class could have a high accuracy due to the fact that the bird is easily distinguished from the background. Theater curtains class could have good results because all the images are characterized by stripes and SIFT takes into account shape information for representing an image, as coral fungus class has a characteristic shape.

As the classes with higher accuracy were studied, the ones with low prediction rate were analyzed too. The worst results, in which no single image of the class was classified correctly (0% ACA), were obtained for the following classes: airedale, chihuahua, great dane and Italian greyhound. Coincidentally, all these classes correspond to dog breeds, which is a classification task that is hard even to humans, due to high variability between individuals.

In order to improve the results color information could be included to this strategy, since PHOW is based on SIFT descriptors that taken into account shape information. To include this information, the same operation can be performed in different color spaces in order to obtain more information that could improve the representation of the images. This color analysis could be improved by taking into account local information by subsampling vicinities around pixels (such as in the pyramid approach). More dimensions could be added to the representation of the image, such as including the relative position of each pixel inside a subpatch. However, care must be taken to avoid overfitting of the model and then increasing the error on the test set. Since the visual dictionary is constructed with k-means clustering algorithm, it could be changed to Gaussian Mixture Model to provide a softer assignment and let the clusters to have shapes different from spherical.

5. Conclusions

- Evaluation of the hyperparameters used in a classification algorithm must be done to choose the best set considering not only ACA result but also time cost required to train the model.

- PHOW demostrated to be a useful algorithm in Caltech 101 dataset but not that much in ImageNet. We presume it is because of the nature of the images on each dataset and the significant difference in the resulting ACA on the test set is also influenced by the fact that ImageNet doubles the number of categories of Caltech-101.

- Some hyperparameters have a big impact on the performance of PHOW algorithm such as number of classes and number of images used in train and test. Spatial X and Y value has a small impact on the performance, whereas others, like number of words or SVM C do not generate a significant change and can lead to overfitting when their values are increased too much.

References

[1] ”Caltech101”, Vision.caltech.edu, 2017. [Online]. Available: `https://www.vision.caltech.edu/Image_Datasets/Caltech101/#Description`. [Accessed: 05- Apr- 2017].

[2] ”ImageNet”, 2017. [Online]. Available: `http://image-net.org/about-overview`

[3] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. IEEE Conference on Computer Vision & Pattern Recognition (CPRV '06), Jun 2006, New York, United States.