Computer Vision Lab 07 -PHOW

Juan Carlos Leon Alcazar
Universidad de los Andes
jc.leon@uniandes.edu.co

1. Description of the datasets

There are tow main image databses used for this laboratory:

Caltech 101: The caltech 101 dataset contains a total of 9145 images distributed in 101 categories. Each category contains between 40 to 800 color images, most of them are catalog images a contain a single instance of a single object. Most of the images are pictures of the object and a few are hand draw. There is some variability in the resolution of images, but most of them have a resolution of around 250x300.

ImageNet: The subset of the image net contains 996 object categories each with 100 instances for a total of 199200 images. All images are in format JPEG contain color information, the image size is standard on the dataset at 250x250. Unlike caltech 101 The categories from imagenet are created according to the WordNet hierarchy.

2. Recognition Method

The base recognition method for the laboratory is the "pyramid histogram of visual words" also know as PHOW. In this method, images are represented by means of an spatial pyramid as proposed proposed by Lazebnik. Such pyramid is built upon a regular grid of N partitions, where N increases(decreases) as the level of the pyramyd increases (decreases). This representation is thus an orderless collection of multi scale feature histograms for a single image.

The features calculated at each region of the grid are the SIFT descriptors proposed by lowe. These features are scale and rotation invariant descriptors created from the gradient information in region of an image. The multi scale features are used to build a codebook or dictionary by using a clustering method (usually Means or Expectation Maximization). This codebook is then used to quantize the descriptor vector for every other image in the train set. Finally, the whole set of quantized vectors is used to train a classifier SVM,

For the classification step features are calculated and quantized on the test set under the same scheme described for the train set, Finally the trained SVM is used for classification.

Changes over the base recognition method
The code base for this laboratory is provided in http://www.vlfeat.org/applications/caltech-101-code.html
As a matlab scripts which uses the VLFeat Library to implement the classification process described above.

2.1. Initial parameter exploration

Before directly using the classifier in the imagenet database, we first try to optimize the hyperparameters in the smaller (and easier caltech) 101 set. The main goal of this procedure is to avoid the hyper parameter search in the imagenet dataset as there is not enough time of complete such task at that scale.

In the base PHOW code we identify 7 Hyperparameters for different stages of the problem [1]

- Number of words for dictionary (any integer number larger than 0)
- Spatial partition in X (any combination of N - $N > 0$ - integers larger than 1)
- Spatial partition in Y (any combination of N - $N > 0$ - integers larger than 1)
- Svm bias multiplier (Any real number)
- Clustering algorithm (Three different approximations to the EM algorithm are available)
- Max number of iterations for the clustering algorithm (Positive Integer)
- Type of kernel (Chi square, intersection kernel, Jensen-Shannon kernel)

[1] The upper limits for the number of iterations in the clustering process and the depth of the KDTree and the image resize operation are also parameters but are considered of far less importance than the mentioned above

- SVM C (Real number)
- SVM solver (sdca, sgd and linear are available)
- Svm Max Number of Iterations (positive integer)
- Epsilon (goodness of fit) of svm. (Real number)

The initial number of possible hyper parameters combinations is not even bounded. To better approach this problem (given the time and computing power limitations), we approach the hyper parameter search as a ..... process first we focus on optimizing a single hyper parameter at the time and the proceed to optimize a second hyperparameter. The full set of hyper parameter is not used, we select a subset of those who could produce larger improvements in the classification process

Hierarchy:

1. Number of words (Changes the number of visual words in the codebook, which are then used to represent every word in the training and test set. This parameter essentially tunes the initial data representation)
2. Spatial partitions (Like the number of words, this parameters changes size of the visual words in the codebook,Again this parameter tunes the initial data representation)
3. SVM C (Controls the loss function of the svm, and thus the global optimization process, by assigning a penalty to the misclassified elements)
4. Kernel (specifies a N-dimensional space where the linear separation of the data is performed, a space where the data is approximately linearly separable will bring the best results)
5. Solver (The algorithm used to linearly separate the data of the SVM)
6. Clustering algorithm (create the codebook from the initial representation of the training images, like the number of words has a great influence in the initial data representation)

The initial parameter tuning was performed with a subset of from caltech (30 images - 15 test, 15 train-) per class with 30 classes, the this allows for a run time of ??? mintes per ters (?? total)

The best set of hyper parameters is:

- Number of words: 1000
- Spatial partitions: [2,3,4]
- SVM C: 50

Clases	Accuracy
10	44
50	21.20
100	13
200	13
500	10.08

Table 1. Acurracy according to the number of clases used for training and test. No test with the complete dataset are performed as the kaget test (500) classes already consumes over 100GB of RAM memory.

- Kernel: Intersection Kernel
- Solver: sdca
- Clustering algorithm: ANN

The performance of this set of parameters over the subset of 30 classes is 80.67%, The performance over the whole set of 100 classes and 9145 Images degrades significantly to is 68%. This figure is only slightly better than the initial configuration at 67.50%.

2.2. Results on Imagenet

Producing a result for the whole dataset is a challenging task, it is not possible to train with the whole 99600 set of training images, it was empirically established that the RAM memory on the server can handle the whole classification process with up to 10000 images (training). This means that either a subset of the images must be selected or a subset of the categories must be selected, we do both;

First we select 996 categories and 10 images each (12.80) Second we select 100 categories and 100 images each (19.84) Third we use an intermediate result with 31 images and 310 images each (14.86)

2.3. Sensitivity to Parameters

This initial configuration of hyper parameters is then used applied in the imagent Train dataset to analyse the response of the method to the following parameters: Number of Categories (50,100,200,500,1000) Size of training set (10,20,50,100) Number of spatial partitions ([2 4] [2 3 4] [2 3 4 5] [2 4 8] [2 4 16]).

Sensitivity to parameter Number of categories 20 training images per class, 2,3,4 spatial partition for both axes.

Sensitivity to parameter Number of categories 50 categories, 2,3,4 spatial partition for both axes.

Sensitivity to parameter Spatial partitions 20 training images per class, 50 categories.

Sensitivity to parameter Number of categories 50 categories, 2,3,4 spatial partition for both axes.

Number of Images	Accuracy
5	25.10
10	36.00
20	45.00
50	52.20
100	56.00

Table 2. Acurracy according to the number of images per class used for training and test.

Spatial Partition	Accuracy
2 4	25.10
2 3 4	36.00
2 3 4 5	45.00
2 4 8	52.20
2 4 8 16	56.00

Table 3. Acurracy according to the spatial partition per image used for training and test.

3. Analysis

Clearly the method is not good enough on the problem of classification on the imagenet dataset, the overall accuracy is under 20

Overall the method is insensitive to the spatial partitions and the initially suggested pyramyd of 2 and 4 paritiions has the same acuracio of deeper pyramys. This suggest that the relevat information can be captured at thes first two scale, and the adtional data of other scales is likely noisy data discarded by the classifier or the clustering process.

It is clear that the classification improves the larger the training set is. This could be attributed to the large amount of different samples from each class which contribute to build a better model of the object (provided the scales and number of classes remains stable).

It is interesting to notice that the smaller the amount of classes in the classification process the better the results. This suggest that either there is not enough capacity in the model to capture the complete variability of a large number of classes, or the base descriptor (Sift) is not capable of capturing the relevant information when the classification problem grows large.

Finally it is also worth noticing that the procedures requires a very large amount of computer resources, the subset used for the laboratory is an order of magnitude smaller than the complete imagenet dataset. Applying this method to the full dataset will probably require require a very large computer or cluster with Terabytes of RAM memory available. Even with today's hardware such requirements are not easy to meet.

3.1. Limitations

The 2 most clear limitations for this approach have been outlined: Model capacity Base descriptors limitations

To address the first issue it would be necessary to change the kernel for the svm, this process is, very difficult as there is not an straightforward method to optimize the representation space, rather this process is mostly (to the authors best knowledge) a trial and error approach. Given the large size of the problem optimizing the kernel could be a very long task, whose results are not granted. Probably the best improvements would be to change the classifier for one that can better handle the large variability in the dataset, it is possible that a very large random forest could perform better in this scenario

Regarding the second limitation, the SIFT descriptors works at the local level, thus, ignoring regional and global patterns in the image that might be critical for the process of classification. A better approach might also use regional information (like the deformable part models) or global information as the second order statistics.

Overall, It is hard to address any of the problems only with modifications to the parameters of the model.

References