Translated from French to English - www.onlinedoctranslator.com

EU Report Introduction to Apprenticeship Automatique

Riazi Ibrahim Luminy Computer License

Ait Kettout Younes Luminy computer license

Zidani Fahed Imed Luminy computer license

The Avengers

Best estimated score by validation crossed:**88%**

Best score obtained on test images mid-term:**83%**

Table of contents

1. Introduction

This project focused on the classification of images that indicates the presence or absence of the sea in a given input image, based on different classifiers and a data-set that contains 414 elements (images). The objective is to have the best possible score on a sample of test images.

In order to meet this expectation, it will first be necessary to choose the best image representation, then the classifier which will give the best score for the test.

Our team has seen itself in difficulty on several occasions, in particular on the representation of the data, an important step before the transition to the selection of the classifier, each member of the Fahd, Ibrahim, Younes group dispersed on its side to collect the maximum of information and sources in order to meet at the end of the week in a two-hour meeting session on discord to discuss ideas, scores obtained and to set the next tasks for the week in the future.

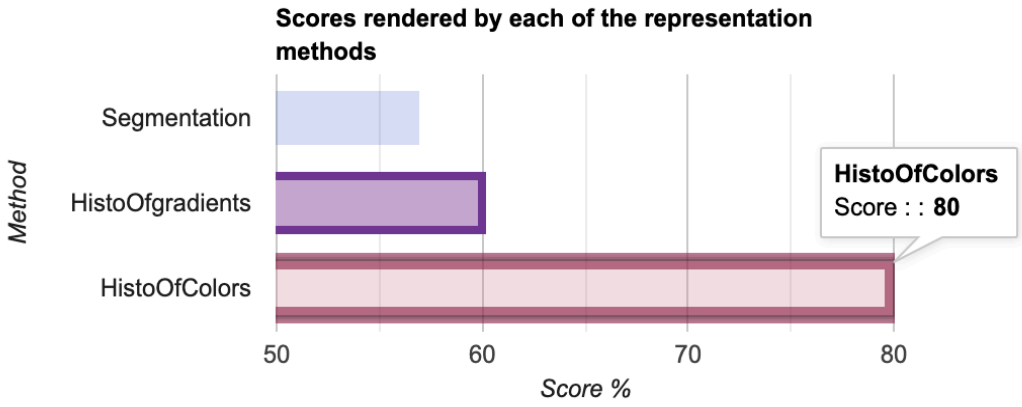
2. Data pre-processing

2.1. Representation of data:

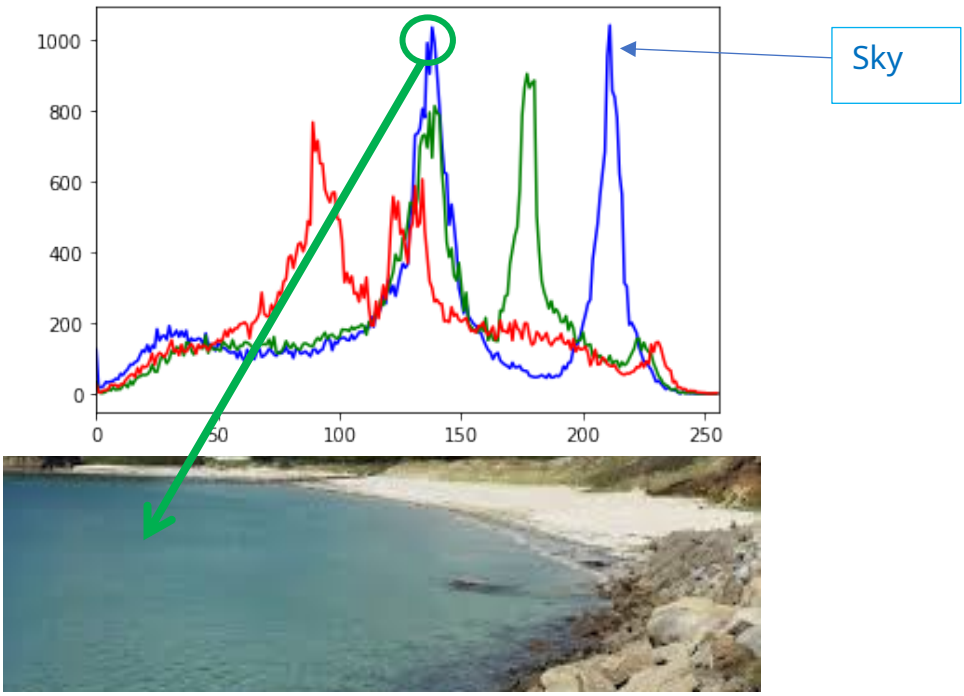
The resolution of the problem of the classification of images according to the presence or not of the sea, amounts to studying the colors constituting the image.

Based on a calculation of the level of the color blue in the images, the presence of the sea in an image means a high level of blue, and its absence means a much lower level except for images with a sky or a blue landscape where the levels of blue will differ.

After having tested several representations (histogram of gradients, segmentation of images, ...) and seeing the results we had, the histogram of colors is the mode of description of image data best suited for this problem.



The color histogram describes well the color channels of RGB (Red Green Blue) images. By digging along this track we have had excellent results.



- Blue color level in a sample of our data-set -

To get there, you must first read an image from a directory path, then transform it into a histogram thanks to the 3 channels and the function ***calchist*()** already predefined in the library ***cv2*** of python.

The processes of resizing, resizing and scaling and cropping images have been skipped, as this will lead us to see less significant results.

Other image representations have been used but without result, since they did not give a better result than the histogram, we quote the representation **felt**tion of images in ***nparray.***

2.2. Data increase:

The Avengers team was content with the initial training dataset, with a dataset of ***414 pictures*** [***207*** images with the presence of sea *(Sea)* and ***207*** images with absence of the sea *(Somewhere else)*] .

Considering it sufficient enough for the resolution of the problem.

Wed: Folder in the project ***Data/Wed*** containing the images with the presence of the sea.

Elsewhere: Folder in the project ***Data/Elsewhere*** containing the images with absence of the sea.

3. Learning algorithm(s) considered

3.1. Adopted algorithm:

The retained algorithm is ***Bagging Bootstrap*** implemented our way.

This choice is due to the excellent result obtained from ***88%*** superior to all the other classifier results, the latter is more suitable for our data-set where it concatenates the result of several classifiers, which have a prediction rate on our test data of *[70% - 79%].*

3.2. Explanation of the selected algorithm(s):

Our algorithm of *Bagging Bootstrapp* is based on the following algorithms:

- ***Logistic regression***.
- ***Bayes naive classifier***.
- ***Vector Machine support (Broad margin separators)***.
- ***Decision trees***

Where each classifier of these algorithms trains on a different sample of data, then each of the models predicts the sample S, and with a majority vote on each entry of the sample it is classified.

Parameterization has not been specified on all classifiers except the **SVC** and **logistic regression** where we specify that they are linear.

The classifiers on which our Bagging algorithm is based have prediction rates of *80% at most* and that even with the changes of the parameters, this absolutely does not call into doubt the power of its algorithms but rather the size of our data-set, where it is difficult to decide for a single classifier and to say that it is the most appropriate for this problem.

Naive Bayes Classifier (naive bayes classifier):

Naive Bayes classifiers are based on Bayes' theorem. One hypothesis retained is that of strong independence hypotheses between the characteristics. These classifiers assume that the value of a particular feature is independent of the value of any other feature.

In a supervised learning situation, Naive Bayes classifiers are trained very effectively. Naive Bayes classifiers need small training data to estimate the parameters needed for classification.

Naive Bayes classifiers are simple in design and implementation and can be applied to many real-world situations.

Logistic regression:

Logistic regression is a method of statistical analysis that involves predicting a data value based on actual observations in a data set.

Logistic regression has become an important tool in the discipline of machine learning.

This approach allows the use of an algorithm in the machine learning application to classify incoming data based on historical data. The more relevant input data there is, the better the algorithm is able to predict classifications within datasets.

Logistic regression is one of the most commonly used machine learning algorithms for binary classification problems, which have two values per class, including predictions such as "this or that", "yes or no" and "A or B".

The purpose of logistic regression is to estimate the probabilities of events and to determine a relationship between characteristics and the probabilities of particular outcomes.

- Vector Machine support (Broad margin separators):

Wide Margin Separators(SVM) are a set of supervised learning methods used for classification, regression, and outlier detection.

The goal of the SVM algorithm is to create the best decision line or boundary that can separate the n-dimensional space into classes so that we can easily place the new data point in the right category in the future. This best decision boundary is called a hyperplane.

SVM picks the extreme points/vectors that help create the hyperplane. These extreme cases are called support vectors and hence the algorithm is called a support vector machine.

The advantages of support vector machines are:

- Effective in large spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of learning points in the decision function (called support vectors), so it is also memory efficient.

- Versatile: Different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. Support vector machines (SVM) are a set of supervised learning methods used for classification, regression, and outlier detection.

Decision trees:

Decision tree is a supervised learning technique that can be used for both classification and regression problems, but it is generally preferred for solving classification problems.

It is a tree classifier, where internal nodes represent features of a dataset, branches represent decision rules, and each leaf node represents the result.

In a decision tree, there are two nodes, which are the decision node and the leaf node. Decision nodes are used to make any decision and have multiple branches while leaf nodes are the output of those decisions and do not contain any other branches.

The decisions or the test are made based on the characteristics of the given data set.

It is a graphical representation of all possible solutions to a problem/ decision given given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which grows on other branches and builds a tree structure.

To build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question and, depending on the answer (Yes/No), it then divides the tree into sub-trees.

4. Classifier performance evaluation

4.1. Performance estimation protocol:

The performance criteria were chosen according to the fact that the different representations of the images gave different scores and these scores differ from one classifier to another.

The estimates were obtained by crossing the results of several classifiers, which have a prediction rate on our test data, varying between [70% - 79%]

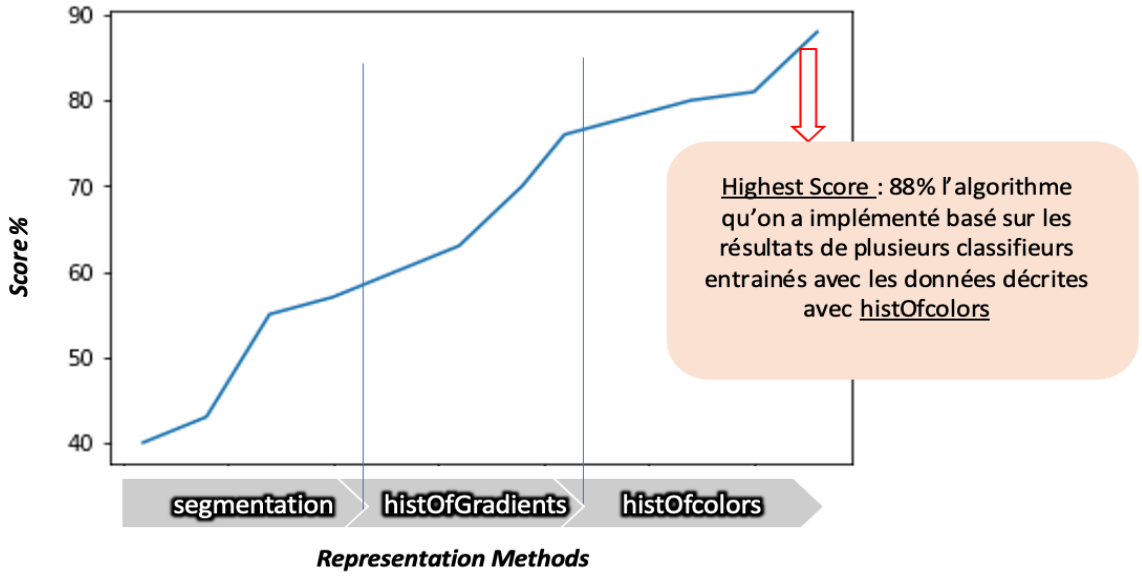
4.2. Performances obtained:

With regard to the protocol that we explained in the previous section, the performances that we obtained by the best approach that we could implement, that of bagging, took us to a score of *88%.*

Knowing that on the mid-term result we had a score of *79% to 80%* .

- The difference is due to the fact that we cross the results of all the classifiers that we test our data-set with, and we make a total score but on the last algorithm that we implemented have crossed the results of all the classifiers which give a note between *70%* and *79%* . This allowed us to considerably increase the score a *88%.*

4.3. Possible performance curves:



5. Conclusion:

This project turned out to be very rewarding as it consisted of a hands-on approach to discovering machine learning through a binary supervised classification design experience.

Despite the working time which was very tight and the small data-set we had available, we were able to achieve a classification giving a high rate.
Taking initiative, respecting deadlines and teamwork were essential aspects of our work.

The main problems we encountered concerned the representation of the data, and the implementation of the latter according to the type accepted by each of the classifiers.

However, if such a classifier were to be implemented perfectly over a period of one month more full-time, it would be necessary to collect more data and adapt a better representation strategy.

References :

- https://iq.opengenus.org/gaussian-naivebayes/#:~:text=Gaussian%20Naive%20Bayes%20supports%20continuous,(independent%20dimensions)%20between%20dimensions .

- https://whatis.techtarget.com/fr/definition/Regressionlogistique#:~:text=La%20r%C3%A9gression%20logistique%20est%20l,et%20%22A%20ou%20B%22.

- https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

- https://scikit-learn.org/stable/modules/svm.html

- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm