Algerian Democratic and Popular Republic
Ministry of Higher Education and Scientific Research

KASDI MERBAH UNIVERSITY OUARGLA
Faculty of New Information and Communication Technologies
Department of Computer Science and Information Technology

License thesis

Computer science

Presented by: Bougoffa Asma Zahrat Arabie & Melouah Messaouda & Guerfi Sahla

Theme

Color image quantization using K-means

Under Supervision: Dr.Mihoub Mazouz
Academic year: 2020/2021

Acknowledgment

First of all, we thank our Almighty God who helped us and gave us the patience and courage during our years of study and to achieve this work.
First and foremost, we would like to thank God Almighty, who throughout years of education has given us health, courage and patience to this day. With great pleasure and deep sense of gratitude, we express our appreciation and sincere thanks to our supervisor, Mazouz Mihoub for his supervision, continuous support and advices provided in an effective manner throughout the completion of this work. We also thank jury members for accepting the chairmanship of the committee and discussing this thesis.

Dedication

Knowing that all words cannot express gratitude, love, respect. I dedicate this humble work
To my dear mother, not enough words can describe how grateful I am to be your daughter. you are my symbol of hope and courage.
To my dear father, you are my role modal. you taught me how to live a life that i am proud of. thank you my hero.
To my dear sisters, Souhir, Bassma and Hadil, thank you for being the best sisters anyone could ever hope for. You always draw a smile on my face.
To my dear brothers, Baghdad, Boualam and Mohamed. Thank you for being a modal brothers who always make feel safe and never in need of anything.
To my teacher Bakhti, thank you so much for everything, for your encouragement, support, kindness. You have been more than just a teacher.

Messaouda Melouah

With the help of Almighty God, who gave me the strength and the patience to be able to carry out this work which I dedicate
To my beloved parents especially my beautiful mother, who is always by my side.
To my sisters, Fatima, Douaa, Maria and Aicha, you are the best sisters in the world.
To my teammates, Messaouda and Sahla, thank you for being understanding and supportive, and I want to thank all my friends (Houaria, Messaouda, Abdellhak).
Special thanks to my grand mother you have a kind heart, my dear aunt Warda thanks for your support and advices, my wonderful uncle Souhaib you have always been the brother, the uncle and best friend.

Bogouffa Asma

With God's luck, I have completed this work with my hard-working colleagues, Massaouda and Asma.
I dedicate this work to my dear mother Dalila and the respected father Amara, may God preserve them and grant me success in their obedience.
I also dedicate this work to the brothers and sisters: Abdullah, Nada, Aicha, Anas, Mouhammed, Maryam.And to all friends from the social and academic environment.

Sahla Guerfi

ملخص

تكميم ألوان الصور تقنية تهدف إلى خفض عدد الألوان المستخدمة لعرض صورة ما على آلة. في هذا العمل سنعرض تطبيقنا للخوارزمية k-means لتنفيذ تقنية تكميم الألوان. تعتبر k-means من خوارزميات تعلم الآلة دون إشراف. تقوم هذه الخوارزمية بتشكيل k مجموعة (clusters) تضم كل منها النقاط الأكثر تجانسا -مقارنة مع نقاط المجموعات الأخرى- و هذا اعتمادا على المسافة الإقليدية بينها. بعد تحميل الصورة و تحديد عدد الألوان (قيمة k)، تقوم الأداة التي طورناها باستخدام لغة البايثون بتطبيق خوارزمية k-means و إنتاج نسخة عن الصورة الابتدائية معروضة فقط بـ "k" لون.

الكلمات المفتاحية: تكميم ألوان الصور، تعلم الآلة، خوار زمية التجميع k-means

ABSTRACT

Image color quantization is a compression technique that aims at reducing the number of colors used to represent an image on a machine. In this work, we will present our application of the K-means algorithm on the color quantization problem. K-means is an unsupervised machine learning algorithm for clustering. The algorithm will form "k" classes (clusters) containing each of them the most homogeneous pixels (with respect to the others belonging to the other clusters) based on the Euclidean distance between them. After loading an image, choosing the number of colors (value of "k"), the tool we have developed in Python, will apply the k-means algorithm and produce another version of the initial image represented only by "k" colors.

Keywords: Image color quantization, K-means clustering algorithm, machine learning.

Résumé

La quantification des couleurs des images est une technique de compression qui vise á diminuer le nombre de couleurs utilisées pour représenter une image sur une machine. Dans ce travail, on va présenter notre application de l'algorithme K-means sur le probléme de quantification des couleurs d'images. K-means est un algorithme d'apprentissage automatique non-supervisé de regroupement (clustering). L'algorithme va former "k" classes (clusters) contenant chacune les pixels les plus homogénes (par rapport aux autres appartenant aux autres clusters) en se basant sur la distance euclidienne entre eux. Aprés avoir chargé une image, choisi le nombre de couleurs (valeur de "k"), l'outil qu'on a développé en Python, va appliquer l'algorithme k-means et produit une autre version de l'image initiale représentée seulement par "k" couleurs.
Mots-clés: Quantification des couleurs des images, Algorithme de regroupement K-means, apprentissage automatique.

Contents

List of Figures

General Introduction

Machine learning is considered one of the most important branches of artificial intelligence. Instead of restricting it to fixed instructions and programming codes, we build models of inputs in the form of algorithms. It aims at making machines able to make decisions in many areas and also came to make predictions, recognize faces, recognize handwriting and many other tasks without being explicitly programmed. These tasks are possible by applying machine learning algorithms which it can be divided into three categories: Supervised, unsupervised and reinforcement learning algorithms. In supervised learning algorithms are based on telling a machine how to relate inputs to outputs by giving them known inputs and their corresponding correct outputs (training dataset) so that it can predict the outputs of new inputs. Unsupervised learning gives machines the ability to approach problems with little or no idea what our results should look like, that is to say, correct answers are note given to it. Reinforcement Learning algorithms refers to the set of methods that allow an agent to learn to choose which action to take, and this independently.

One of the problems we face in our days is that of immensity of data. This will continue to increase, and no matter how high the storage space is. To preserve some storage space, some solutions have been introduced as lossless compression which is commonly required for textual content and facts files and lossy compression for compressing multimedia (audio, image). Color image quantization, is one of the lossy compression techniques. quantizing the colors of an image is a technique that aims to reduce the number of colors needed to represent it. This is very important, for example, to represent an image on devices that only support a limited number of colors. The process of reducing the number of colors to k requires forming k classes of pixels containing each of them the most homogeneous pixels relatively to others.

The most well-known clustering algorithm in machine learning is K-means. It is an unsupervised algorithm which offers a solution to the clustering problem. In short, it aims at forming k classes (k known in advance) containing homogeneous data points by following well-defined steps. In this regard, we applied the K-means algorithm to quantize the colors of a given image by implementing a python-based tool. The rest of this manuscript is divided into three chapters as follows:

Chapter 01| Machine learning In this chapter, a brief overview on machine learning field is given in terms of its definition, its algorithms and its application.

Chapter 02| K-means In this chapter, a detailed description of the well-known clustering algorithm, K-means, is given.

Chapter 03 | Image quantization using K-means The last chapter is devoted to the application of k-means algorithm for image quantization where the developed tool is represent.

CHAPTER 1

MACHINE LEARNING

1.1 Introduction

Now humans are able to teach machines how to make decisions in many areas and also to make predictions, recognize faces, recognize handwriting, and many other tasks that we thought it was impossible for machines to do. Machine learning is considered one of the most important branches of artificial intelligence. Instead of restricting it to fixed instructions and programming codes, we build models of inputs in the form of algorithms. Arthur Samuel described Machine Learning as the:"field of study that gives computer the ability to learn without being explicitly programmed"[7] How was the impact of machine learning on various areas of life?

1.2 Application Of Machine Learning

Starting with our simple daily life : machine learning is used to automatically classify our email into spam or important email by analyzing the data contained in the message. Moving to the application of machine learning in the commercial field, where the largest companies depend on it, as it satisfies their customers and increases their profits with a deliberate step, which is to study their desires and needs, thus providing them with what is required, She even replaced the customer service staff with chatbots and used the employee only in case the request was complicated, the chatbots forward the call to him

Other applications :

Web search :
rating web page primarily based totally on what you are maximum probably to click on.
Computational biology :
rational layout pills with inside the PC primarily based totally on experiments. Space exploration: space probes and radio astronomy.

Finance :
determine who to ship what credit scorecard offers to, Evaluation of hazard on credit score offers, How to determine wherein to make investments money.
Robotics :
how to deal with uncertainty in new environments. Autonomous Self-driving cars.
E-commerce :
predicting patron churn. Whether or no longer a transaction is fraudulent.
Social media :
data about relationships and preferences, machine learning extracts value from data.
Debugging :
in arithmetic tasks Time-consuming process, this will suggest where the error might be.
Information extraction :
Ask questions about databases on the internet.

THE LARGEST COMPANIES THAT INVEST IN MACHINE LEARNING :

- Whereas, Facebook has the largest face database in the world, and this is to achieve visual recognition, using the data that is fed by application users.
- Google uses it to rank search engines and image services.
- Apple is also investing significant financial resources in this area for the Siri service that it provides to the user on its mobile devices. And many other areas where machine learning represents the bulk of its work, such as the field of weather forecasting, the stock exchange, and so on.
- IBM company has produced what is called a giant machine learning "WESTON", which is widely used in the financial and medical sectors. Microsoft also provided AZURE ML STUDIO, Also Google Tensor Flow service And many free open source programs that allow the public to learn the machine and develop applications.[5]

1.3 TYPES OF MACHINE LEARNING :

First, there are two levels of machine learning: Inductive Learning : is how AI systems attempt to use a generalized rule to carry out observations Deductive :which applies general provisions in specific examples. Opinion differ about the types of machine learning, So we will mention three types:

1.3.1 SUPERVISED LEARNING

Supervised learning(Predictive learning) : is based on defining a machine how to relate inputs to outputs by training it using known inputs and outputs so that it can predict the outputs of new inputs.[11] There is two tasks of supervised learning: regression and classification

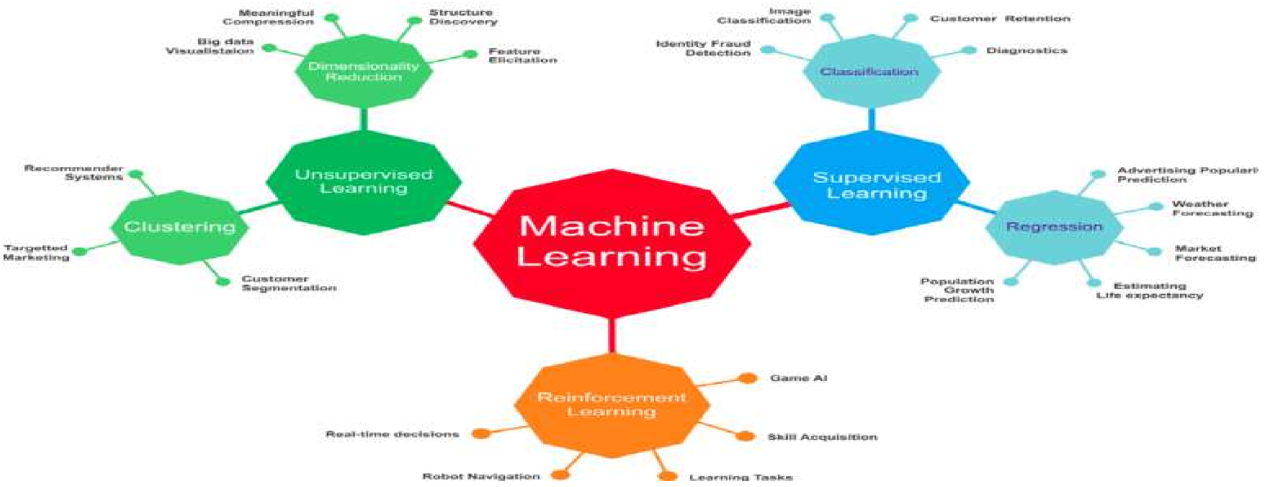


Figure 1.1: Types Of Machine Learning

CLASSIFICATION :

is the process of dividing a specific data set into classes. you can do this for both structured and unstructured data. The process starts by predicting the data category. The class is often mentioned, as a target , label or category.[11]

Classification predictive modeling is the problem of approximating the mapping function (f) of the input variable (x) to the discrete output variable (y). Output variables are usually called labels or categories. The indicator function can predict the category or category of a particular observation.[9] Ex: classification of students according to their rates to fail or pass.

REGRESSION :

this is a mathematical technique in which a data scientist can predict a continuous outcome (y) based on the value of one or more predictor variables(x).[11]

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y), which is a real-value as an integer.[9]
Ex: We predict the house price according to its size after comparing it to the size and prices of other homes.

$$Y = f(x) + \epsilon$$

X (input) = house area
[11]
Y (Output) = Price of the house
f = Function describing the relationship between X and Y
$\epsilon(epsilon) = randomerrorterm(positiveornegative)withmeanzero$

- Regression can include both real and discrete input variables.
- The regression problem in which the input variables are sorted by time is called the time series of the prediction problem.

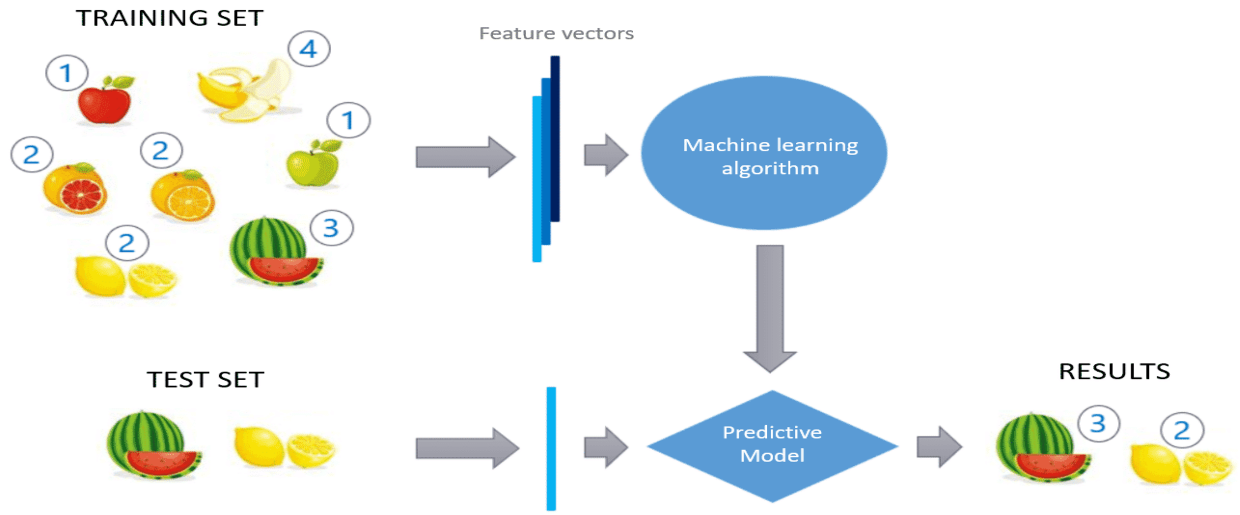


Figure 1.2: picture explain supervised learning

- Problems with multiple input variables are usually called multiple regression problems.

How to avoid errors and to confirm the correctness and results of predictive modeling of the regression ? There are many ways to evaluate the capabilities of regression prediction models. However, the most common is to calculate the mean square error RMSE.[9]

RMSE = sqrt (average (error^2))

Linear regression : Which refers to a regression model composed entirely of linear variables. Starting from the simple case , linear regression with one variable is a technique used to model the relationship between an independent variable with an input (function variable) and an output. Structured variable the usage of a linear version i.e. a line.[11]
Simple linear equation :

$$y = a0 + a1 * x1$$

The most common case is multiple linear regression , which is a model of relationship between several independent input variables (feature variables) and dependent output variables, in the sense that the output is a linear combination of input variables, the model remains linear . We can model multiple linear regression as follows :

$$Y = a1 * X1 + a2 * X2 + a3 * X3 + ... + an * Xn + b$$

a_n: the coefficients
X_n: the variables
b: the bias

as we will see , this characteristic does now no longer encompass any non-linearities and so is handiest appropriate for modeling linearly separable data. it is pretty clean to recognize as we are clearly weighting the significance of every characteristic variable Xn the use of the coefficient weights an .We decide those weights an and the bias b the use of a Stochastic Gradient Descent (SGD: is a technique used to find the model parameters that correspond to the best fit between predicted and actual outputs).[9]

Polynomial Regression: is a regression algorithm that models the relationship between the dependent variable(y) and independent variable(x) as a polynomial of degree n.[9] If you want to build a model suitable for processing nonlinear shared data , you need to use polynomial regression .Using this regression method ,the best fit line is not a straight line , but a curve that fits the data points. In polynomial regression the cardinality of some independent variables is greater than 1.

The polynomial regression equation:

$$Y = b0 + b1 * X + b2 * X^2 + ... + bn * X^n$$

Y : dependent variable

b : coefficients

X : independent variable

- Polynomial regression requires carful design.
- Full control over the modeling of feature variables(exponent to set).
- Need some knowledge of the data in order to select the best exponents.

1.3.2 Unsupervised Learning :

Unsupervised Learning: is a machine learning technique in which users do not have to follow the model , but the model can work on its own to discover patterns and information that have not discovered before, it mainly deals with the unlabeled data.[10]

- unsupervised strategies assist you to discover capabilities which may be beneficial for categorization.
- unsupervised machine learning can finds various unknown patterns in the data.

There are two types of unsupervised learning association and clustering.

Association :

you can use association rules to create mappings between data objects in large databases. This unattended method aims to find interesting relationships between variables in large databases.
Ex: a subgroup of books grouped by date of authorship.

Clustering :

is finding a structures or patterns in a collection of unclassified data. The clustering algorithm processes your data and looks for natural clusters (groups), if these clusters exist in the data. You can also change the number of clusters that the algorithm should recognize. Adjust the granularity of these groups.[10]

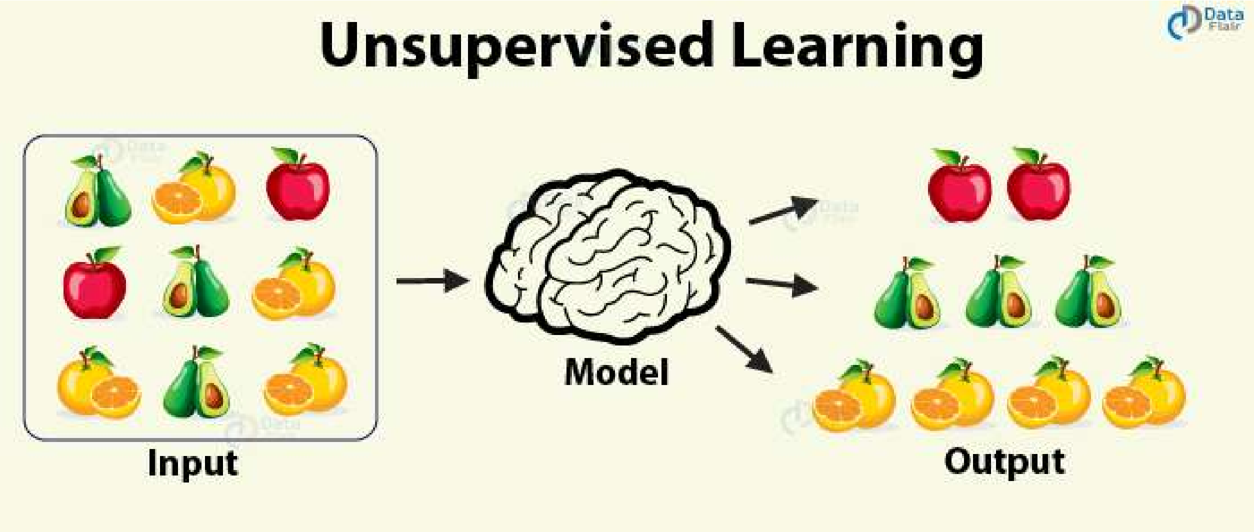


Figure 1.3: Explaining of unsupervised learning

The different types of clustering methods:

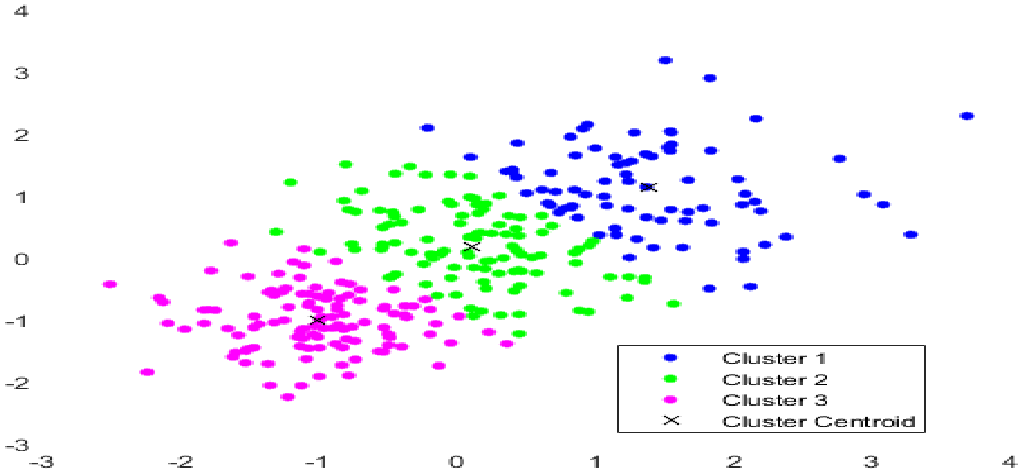


Figure 1.4: picture that explains clustering

EXCLUSIVE :

using this clustering method , data can be grouped so that one data element can only belong to one group.[10]
Exclusive : K-means (this is an iterative grouping algorithm that allows you to find the maximum value for each iteration. First select the required number of groups .Using this clustering method you must group the data points into K groups).[10]
Notice : when K means is greater the groups get smaller with higher accuracy vice versa a lower K means larger groups with less accuracy .

Agglomerative : using this grouping method ,each data item is a grouping .The iterative join between two nearby clusters reduces the number of clusters.[10] **Hierarchical** clustering (is an algorithm for creating a cluster hierarchy .It starts with all the data allocated to a single cluster. In the same cluster ,there are two clusters. When there is only one group left, the algorithm ends) also it is considered the most popular and widely used method to analyze social network data.[10]
Overlapping : this method uses fuzzy sets to group data, each point can belong to two or more groups with different degrees of memberships . Here the data is linked with the corresponding memberships value. **Fuzzy c-means clustering** this algorithm assings membership to each data point corresponding to each center in the group based on the distance between the center of group and the data point.

1.3.3 REINFORCEMENT LEARNING :

is a method that allows to maximize the use of accumulated rewards, which deals with how software agents perform operations in the environment.[5]
Notice: reinforcement learning belong to deep learning . Consistent decision time is essential to reinforcement problems. The actions of the agents determine the subsequent data they receive. There is no supervisor ,only a real number or a reward signal.

There are two kinds of reinforcement learning:

POSITIVE :

It is defined as an event that occurs due to a certain behavior , which increases the intensity and frequency of the behavior, and positively affects the actions taken by the agent.[5] This type of Reinforcement helps you to maximize performance and sustain change for a more extended period. However, too much Reinforcement may lead to over-optimization of state, which can affect the results.

NEGATIVE :

is defined as an increase in behavior due to negative condition that should be stopped or avoided.[5] It helps you to define the minimum stand of performance. However, the drawback of this method is that it provides enough to meet up the minimum behavior.
Important terms in reinforcement algorithms:[8]

- Environment (e) : the scenario that the agent must deal with.
- Reward (R) : A feedback returned to the agent from the environment to evaluate the action of the agent.
- Agent : a potential entity that takes action in the environment to obtain compensation.
- Value (V) : this is the long-term discount period of expected performance relative to short-term rewards.
- State (S): state refers back to the present day state of affairs back with the aid of using the environment .

- Value function: it specifies the cost of a nation this is the overall quantity of reward .it is an agent which have to be predicted starting from that nation.
- Policy (π) : this is the strategy for the agent to decide what to do next based on the current state.

.

REINFORCEMENT LEARNING ALGORITHMS :

Model-Based : in this reinforcement learning method, you want to create a digital version for every environment the agent learns to carry out in that unique environment .[8]
Value-Based : in value based totally reinforcement learning method, you need to try and maximize value function V(s). in this method, the agent is looking ahead to a longâtime period go back of the modern states beneath Neath policy π.[8]
Policy-Based : is to find the optimal policy for the maximum future rewards without using the value function. There are two types of policy (Deterministic, Stochastic).[8]

1.4 ADVANTAGES AND DISADVANTAGES OF MACHINE LEARNING :

ADVANTAGES :

- No human intervention needed(automation): with machine learning ,you don't need to worry that the machine learn by it self.Make predictions and improve every step of the algorithm. Virus software ,once you find new threats, you will learn to filter them out. ML is also good at detecting spam(as we mentioned above).
- Handling multi-dimensional and multi-variety data: machine learning algorithms are very suitable for processing multi-dimensional and diverse data , and can be processed in dynamic or insecure environments.
- Continuous improvement: as ML benefit experience , they hold enhancing in accuracy and efficiency. This allows them to make higher decisions. Say you want to make a climate forecast model.As the quantity of information you've got continues growing, your algorithms discover ways to make extra correct predictions faster.
- Wide applications: you will be an e-trailer or a healthcare company and make ML paintings for you. Where it does apply , it holds the functionality to assist supply a miles greater non-public revel in to clients whilst additionally focused on the proper clients.

DISADVANTAGES :

- Time and resources: Ml desires sufficient time to permit the algorithms research and increase sufficient to meet their cause with large amount of accuracy and relevancy. It additionally desires big assets to function .This can imply extra necessities of laptop energy for you.

- Data acquisition: machine learning calls for big records units to educate on , And those need to be inclusive/unbiased , and of true quality .There also can be instances where in they ought to anticipate new records to be generated.
- Interpretation of results: another essential assignment is the cap potential to correctly interpret effects generated through the algorithms .You have to additionally cautiously, pick out the algorithms to your purpose.[7]

1.5 CONCLUSION :

Machine learning is the core area of artificial intelligence, the computer can enter the self-learning mode without explicit programming after loading new data , these computers will learn, grow, change and develop on their own. ML can be supervised or unsupervised .if your data volume is small and the training data labels are clear, so choose supervised learning, unsupervised learning can usually provide better performance and results for large amounts of data.[7]

CHAPTER 2

K-means

2.1 Introduction

Nowadays, clustering techniques is very common and important, and as the amount of data increases, the number of cluster tends to increase. K-means is a simple clustering analysis technique that aims to find the best way to divide n entities into k groups (so-called clusters). To minimize the total distance between group members and the corresponding center, regardless of the group. Each entity belongs to the cluster with the nearest mean.

K-means clustering is a section-based grouping method used to classify/group items into k groups (where k is the number of user-defined groups). The grouping is done by minimizing the sum of the squared distance[6] (Euclidean distance) between the element and the corresponding centroid.

Although K-means is simple and can be used for multiple data types, it is very sensitive to the starting position of the cluster center. There are two simple ways to initialize the center of the cluster, namely, randomly selecting the seed value or selecting the first k samples of data points. Or, select a different set of starting values (from data points), and then select the data closest to the optimal set. However, testing different initial sets is considered impractical, especially for a large number of clusters. In addition, the original K-Means algorithm is very computationally intensive, especially when dealing with large amounts of data sets.

2.2 Clustering Vs Classification

Clustering and classification techniques are used in machine learning, information retrieval, image exploration and related tasks. These two strategies are the two main parts of the data mining process. In the field of data analysis, they need to be used to enhance algorithms. These processes divide the data into sets. In today's information age, this task is very important, because as development progresses, huge data growth must be appropriately promoted. Interestingly, grouping and ranking can help solve global problems such as crime, poverty and disease through data science.

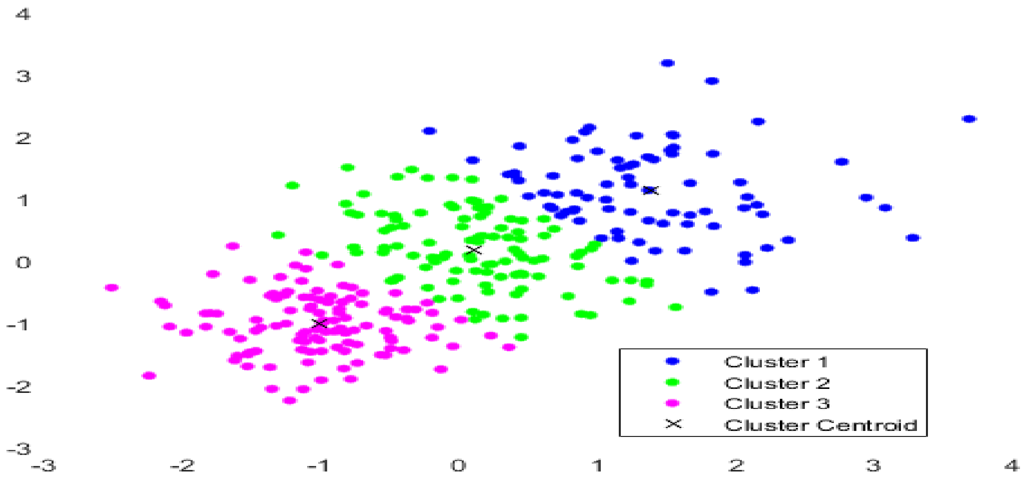


Figure 2.1: demonstration of k-means clustering

- The essential difference is that clustering is unsupervised and is considered "self-learning" while classification is supervised because it is based on predefined labels.

- Clustering does not use training sets, which are groups of instances used to generate the clusters, while classification is a critical step in using training sets to identify similar characteristics.

- Clustering works with unlabeled data because no training is required. On the other hand, classification deals with labeled and unlabeled data in your processes

- Clustering groups objects together with the aim of reducing relationships and learning new information from hidden patterns, while classification tries to determine which explicit group a particular object belongs to.

- While classification does not indicate what needs to be learned, the clustering indicates the improvement needed by pointing out the differences when the similarities between the data are taken into account.

- In general, clustering only consists of a single phase (clustering), while classification consists of two phases: training (the model learns from the training data set) and testing (the target class is predicted).

- Compared to clustering, classification is more related to prediction, as it specifically aims to identify the target classes. This can be used, for example, in the "detection of key points on the face", since it can be used to predict whether a particular witness is lying or not.

- Since the classification is made up of more levels, deals with predictions, and includes grades or levels, its nature is more complicated compared to grouping, which is mainly about grouping similar attributes.

- Clustering algorithms are mainly linear and non-linear, while classification includes more algorithm tools, such as linear classifiers, neural networks, kernel estimation, decision trees and auxiliary vector machines.

2.3 K-MEANS CLUSTERING ALGORITHM

The implementation of the k-means algorithm can divide the data points of the data set into several groups based on the closest average value. The k-means clustering algorithm can be used to determine the best division to divide the data points into groups so that the distance between the points in each group is minimized.

The process follows a simple method of classifying a given data set into a given number of clusters (assuming there are k clusters). The basic idea is to define k centers, one for each cluster. These centers need to be positioned correctly, because different locations will produce different results. Therefore, it is best to keep them as far away from each other as possible. The next step is to associate each point belonging to the specified data set with the nearest center.

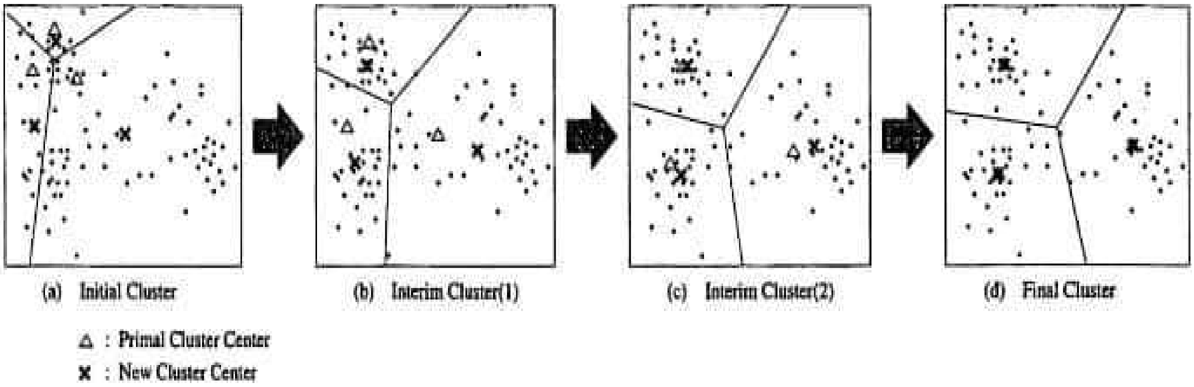
If there is no unfinished data, complete the first step and execute the young group. At this stage, we need to calculate new k centroids as the centroids of the clusters obtained as a result of the previous step. For the new k center of gravity, a simultaneous connection must be established between the same given data point and the new closest center. Create a loop. As a result of this loop, we can see that the k centers gradually change their positions until no more changes are made, in other words, the centers stop moving. Finally, the algorithm aims to minimize the objective function called the quadratic error function.

2.3.1 PROPERTIES OF K-MEANS

- The algorithm try to determine the K parts that minimize the square of the error function.
- This method usually leads to local optimum.
- This is a well-known method based on the center of gravity, which uses the parameter k and divides the set of n objects into k groups, so that the obtained similarity within the group is the same, but the similarity between the groups is low.
- Clustering similarity is measured relative to the average value of objects in the group whose center or center of gravity can be seen.

K-means algorithm can be summarized as follow:[3]

- Choose K initial cluster centers [Fig.2(a)].
- Assign all elements to the cluster nearest the element.
- Make the new cluster centers the center of gravity of the new clusters(Fig. 2).
- If all new cluster centers are the same as the previous cluster centers [such as Fig. 2(d)], the algorithm is finished. If they are not, go back to step (2).



(a) Initial Cluster (b) Interim Cluster(1) (c) Interim Cluster(2) (d) Final Cluster

Figure 2.2: k-means algorithm[6]

2.4 Advantages and disadvantages of K-means

Advantages:

- It is fast, reliable and easy to understand.
- As an unsupervised method and effective enough.
- Flexible and easy to explain
- Suitable for large amounts of data.
- If data sets are different, better results will be obtained.
- Compared with other complex clustering methods, it is easy to implement.
- Produce more cohesive groups.
- When the group is spherical, the K-Means algorithm is suitable for defining the data structure.
- When the centroid is recalculated, the group changes.
- There are no assumptions about the data distribution.
- If the variables are large, then in most cases, if we keep k small, the K-means will be faster than hierarchical grouping.

DISADVANTAGES:

- You need to determine the number of cluster centers in advance.
- It is difficult to predict the value of K.
- If there are two data that overlap greatly, they cannot be distinguished, nor can it be said that there are two clusters.
- It does not apply to global clusters.
- Different data representations will lead to different results.
- Different starting partitions may lead to different ending clusters.
- The Euclidean distance will weight the factors unevenly.
- Sometimes, randomly choosing centroids may not produce fruitful results.
- Not applicable to clusters of different sizes and different size densities (in the original data).
- It can only be used if a value is defined.
- Scaling sensitivity : If a large amount of data is detected, the computer may crash.
- The K-Means algorithm does not allow data points that are far away from each other to belong to the same group, even if they obviously belong to the same cluster.
- If the group has a complex geometric shape, the K-means will not be able to group the data well.

2.5 ENHANCEMENT OF THE K-MEANS :

K-Means algorithm can be anhanced by the following steps:

- 1)Randomly choose one of the observations to be a cluster center.
- 2)For each observation x, determine d(x), where d(x) denotes the MINIMAL distance from x to a current cluster center.
- 3)Choose next cluster center from the data points, with probability of making an observation x a cluster center proportional to d(x).
- 4)Repeat 2 and 3 until you have chosen the right number of clusters.

This process has a setup cost, but convergence tends to be faster and better (lower heterogeneity)

2.6 Conclusion

One of the most famous clustering algorithms is K-means clustering. Generally, when solving clustering problems, professionals first need to understand the structure of the data set. The purpose of k-means is to group data points at subgroups that do not overlap. When the shape of the cluster is slightly spherical, it does a good job, but when the geometry of the cluster deviates from the spherical shape, it is affected. In addition, it does not know the number of clusters in the data, so it needs to be predefined. Understanding the assumptions behind the algorithms and methods is always helpful to understand the advantages and disadvantages of each method. You can decide when and under what circumstances to use each form.

CHAPTER 3

COLOR IMAGE QUANTIZATION USING K-MEANS

3.1 INTRODUCTION

Clustering belongs to the unsupervised machine learning Which has become used in and several fields like marketing and sales, identifying fake news, Many companies use it to collect their data and extract information And as we know, these companies have a huge amount of data. The problematize here is that the data will continue to increase, and no matter how high the storage space is, it will be filled one day. To make that day further away, or rather to preserve some storage space. In this regard, we applied one of the lossy compression techniques which is color image quantization is one of the maximum often used operations in computer graphics, this process was previously used to solve the problem of video graphics adapters where many bytes per pixel cannot display which could only support 8 bites color per pixel max, as result we were able to display a 24 bites digital color image on these devices. And today, almost all graphics hardware is 24 bites, So we will use image quantization for another goal, which is to reduce the space required to store images this is done using one of the clustering algorithms.[1]

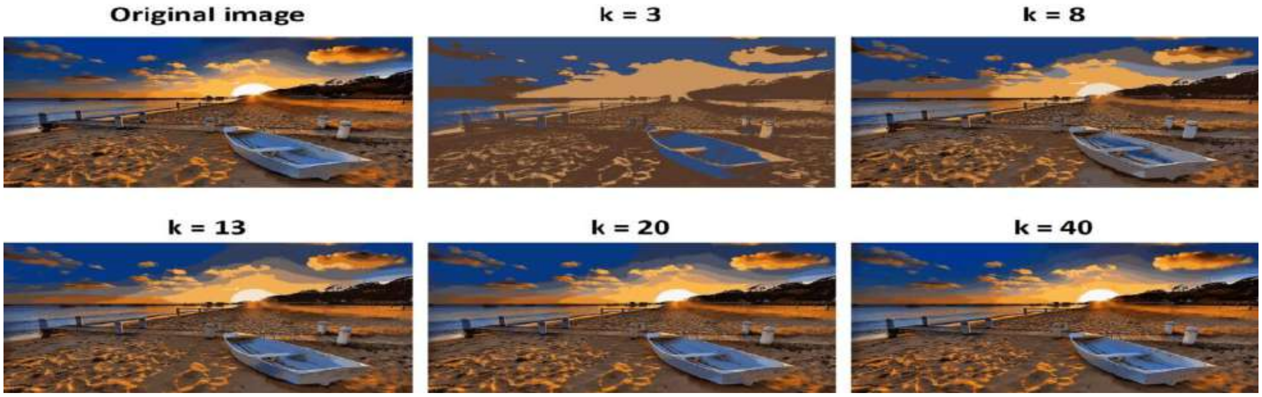


Figure 3.1: demonstration of color quantization using clustering

3.2 COLOR IMAGE QUANTIZATION

In quantization, the representation value is assigned to the range of input values. In image processing, the quantized value can be analog or digital. In color image quantization, a set of colors is selected to represent the color gamut of the image. And calculate the distribution of the color space to the representative color. There are two main types of quantification methods: uniform and conical. Through uniform quantization, the range of input variables is divided into equally long intervals. Select the interval of the cone quantization. It is usually based on the statistical distribution of the input variables. In order to compare the quality of different assessments, deviation measures or error measures are usually introduced. With this formalism, you can search for the "best" conical quantification of a variable (or image).[2]

3.3 PYTHON OVERVIEW

Programmers prefer Python because it is fast and easy to use. Python cuts development time in half thanks to its easy-to-read syntax and simple assembly functions. Thanks to the built-in debugger, debugging in Python is very easy. Most programmers are forced to use Python. Python remains the first choice for data scientists who use it to create and use machine learning and other scientific computing applications. Python can run on Windows, Linux/Unix and Mac OS, and is suitable for Java and .NET virtual machines. Thanks to an open source license vetted by the OSI, Python can also be used in commercial products for free. Python has become the language of choice for data analysis, and the growing research trend in Python is also showing that Python is the next "big thing" and a must-have for data science professionals.

3.4 AN APPLICATION FOR COLOR IMAGE QUANTIZATION USING PYTHON

3.4.1 CLASS DIAGRAM :

There are two classes in this diagram which are:
class cluster and class k-means, where Each of these classes has its own methods and attributes.
Cluster attributes(privet): Pixels, Image and Centroid.
method (public): def moyenne to calculate new centroid.
k-means attributes (public): pexls, clusters, K, image.
methods(public): def fit image for read img, cal-dis forr calculate the distance Between the clausters and the centeroid, min-ind for find the minimum distance.
As we can see the cluster belongs to a single class k-means, but the class k-means can creat many k-means clusters.

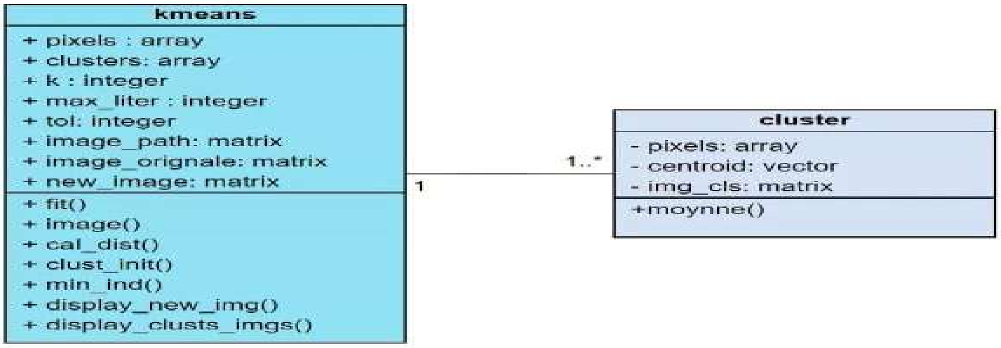


Figure 3.2: application class diagram

3.4.2 IMPLEMENTATION

LIBRARIES:

OpenCV OpenCV was started at Intel in 1999 by Gary Bradsky and the first release came out in 2000. Vadim Pisarevsky joined Gary Bradsky to manage Intelâs Russian software OpenCV team . It is a Python library[4], open source machine learning and machine vision software. The library contains a complete set of robust algorithms. It supports multiple programming languages and can run on multiple operating system platforms.

NumPy NumPy, short for Numerical Python, is the foundational package for scientific computing in Python[16]. It consists of pointers to memory and metadata. Metadata is especially used to interpret the data stored in it. "Data Type", "Table" and "Step". In addition to the fast array processing capabilities that NumPy adds to Python, one of the main goals of data analysis is the main container for passing data between algorithms. For numerical data, compared with other built-in Python data structures, NumPy arrays are a more effective way of data storage and management. In addition, libraries written in low-level languages such as C or Fortran can handle data stored in Numpy arrays without copying the data.

PIL The Python image library (PIL) is a free and open source supporting library of the Python programming language, which supports opening, editing and saving a variety of different image file formats. It is available for Windows, Mac OS X and Linux. The latest version of PIL is 1.7, which was released in September 2009, supports Python 3, and will be released "later". The development of the original project (called PIL) ceased in 2011. Subsequently, a follow-up project called Pillow was separated from the PIL repository and added support for Python 3. This fork was adopted as an alternative to the original PIL in Linux distributions. Pillow provides several standard image processing methods, including: pixel processing, masking and transparency processing, image filtering (such as blur, contour, anti-aliasing or edge detection), image enhancement (such as sharpness, brightness, contrast, etc.) Adjust colors, add text to the image, and so on.

Matplotlib matplotlib is the most popular Python library for graphing and other visualizations of 2D data. It was originally created by John D. Hunter (JDH) and is now maintained by an excellent development team. It is ideal for creating publishable charts. It is well integrated with IPython (see below) and provides a convenient interactive environment for plotting and exploring data. The graph is also interactive; you can zoom in on a part of the graph and scroll it using the graph window toolbar.

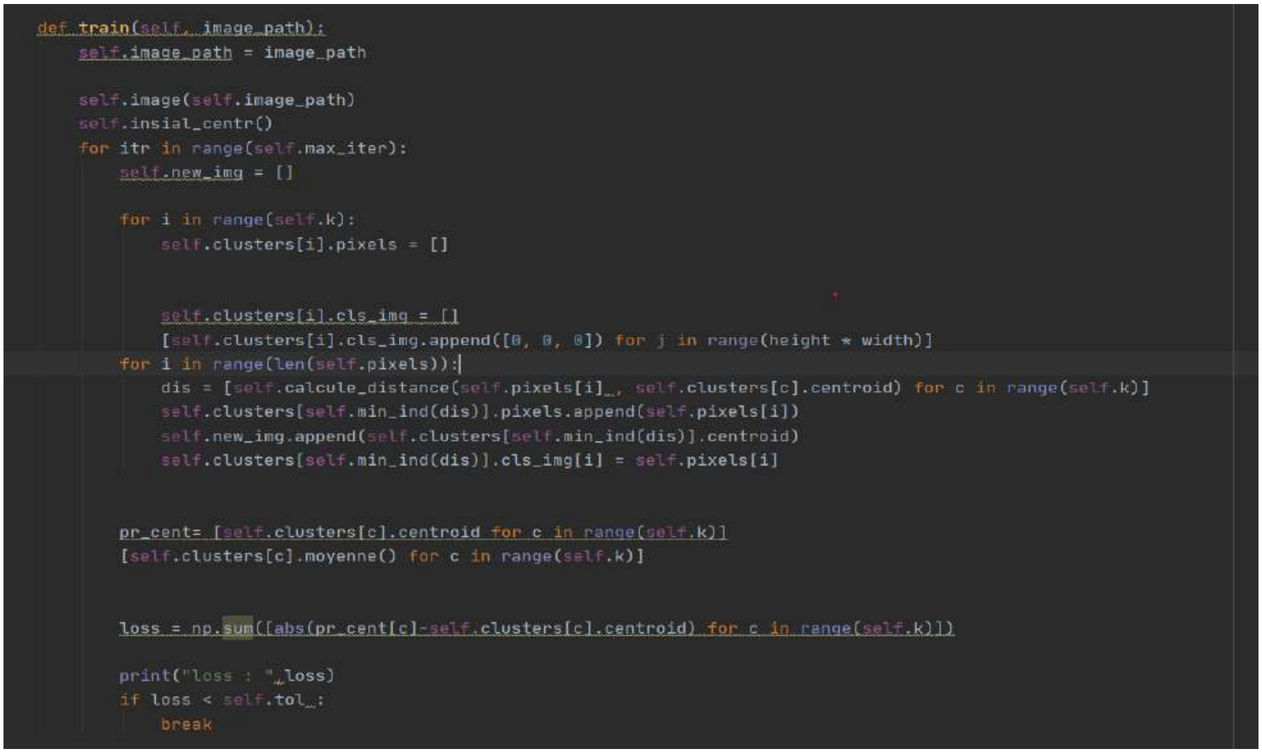
3.4.3 MAIN METHODS

we built a desktop application with a friendly user interface That uses K-Means in order to perform color Quantization on a chosen image by the user, by applying clustering on the pixels of that image.

The user has the ability to select a picture of his desire by pressing the upload button. Besides, he can also choose the number k (the number of clusters) by using the simple to use spin box. Then, by pressing the run button, the application would apply clustering on the pixels and display the result in a form of an optimized new picture. The selected image is read as soon as we know the schema of the image. Then, we will extract the pixel set of the image and put it in a list. Afterwords, we choose the K centroids at random. Withing each iteration, every pixel would be assigned to its cluster, based on measuring the distance between the pixel and every centroid then selecting the closest to that pixel. Each cluster has its own image at the beginning of each iteration the cluster images would be empty, then every pixel is added to the image of the cluster that is allocated to. And the pixels would be replaced and added in a new image that represents the color quantized image. After that, we store the centroid then update them by calculating the average of the cluster points. And since the centroids and their distance between them and the pixels would change, Therefore, the clusters list of pixels and image would be erased. These steps would loop until the convergence. At the end of each epoch, we sum the subtraction of the old and new centroid to compute the changes of the clustering assignments, of the changes is less than a given tolerance number, the training process would stop. Finally, the application would display the quantized image and save each cluster image locally.

Aside benefit of our approach is that the cluster images can be used for segmentation. This can be particularly helpful when dealing with data that are hard to label. Such as medical scans and plant leaf disease. Our work deals with the lack of an application for color quantization, especially using python language. For this purpose, we build a friendly-to-use interface with the PyQt5 python library.

Furthermore, we aimed to make a unique k-means structure by dividing the algorithm into two classes instead of one, which are cluster and k-means classes. This has proven to be more readable and easier to understand.



```
def train(self, image_path):
    self.image_path = image_path

    self.image(self.image_path)
    self.insial_centr()
    for itr in range(self.max_iter):
        self.new_img = []

        for i in range(self.k):
            self.clusters[i].pixels = []


            self.clusters[i].cls_img = []
            [self.clusters[i].cls_img.append([0, 0, 0]) for j in range(height * width)]
        for i in range(len(self.pixels)):
            dis = [self.calcule_distance(self.pixels[i] , self.clusters[c].centroid) for c in range(self.k)]
            self.clusters[self.min_ind(dis)].pixels.append(self.pixels[i])
            self.new_img.append(self.clusters[self.min_ind(dis)].centroid)
            self.clusters[self.min_ind(dis)].cls_img[i] = self.pixels[i]


        pr_cent= [self.clusters[c].centroid for c in range(self.k)]
        [self.clusters[c].moyenne() for c in range(self.k)]


        loss = np.sum([abs(pr_cent[c]-self.clusters[c].centroid) for c in range(self.k)])

        print("loss : " ,loss)
        if loss < self.tol :
            break
```

Figure 3.3: a capture of our main k-means training process

3.4.4 OBTAINED RESULTS

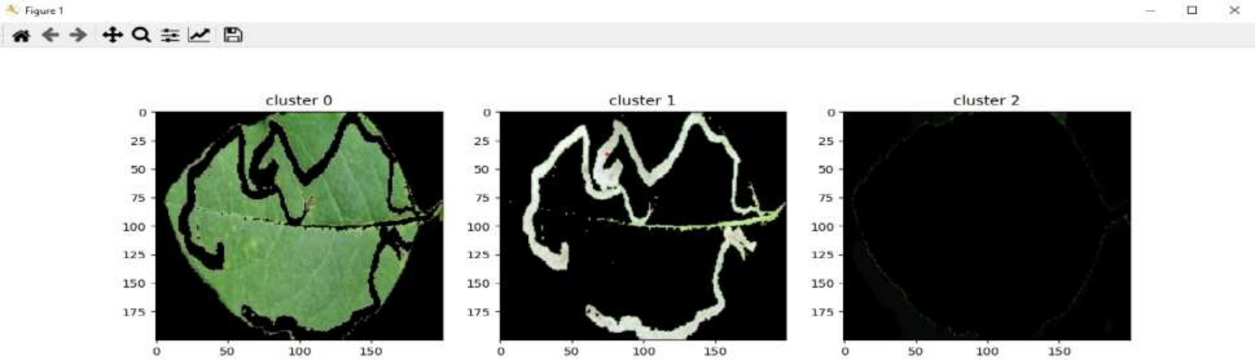


Figure 3.4: a caption of each cluster image

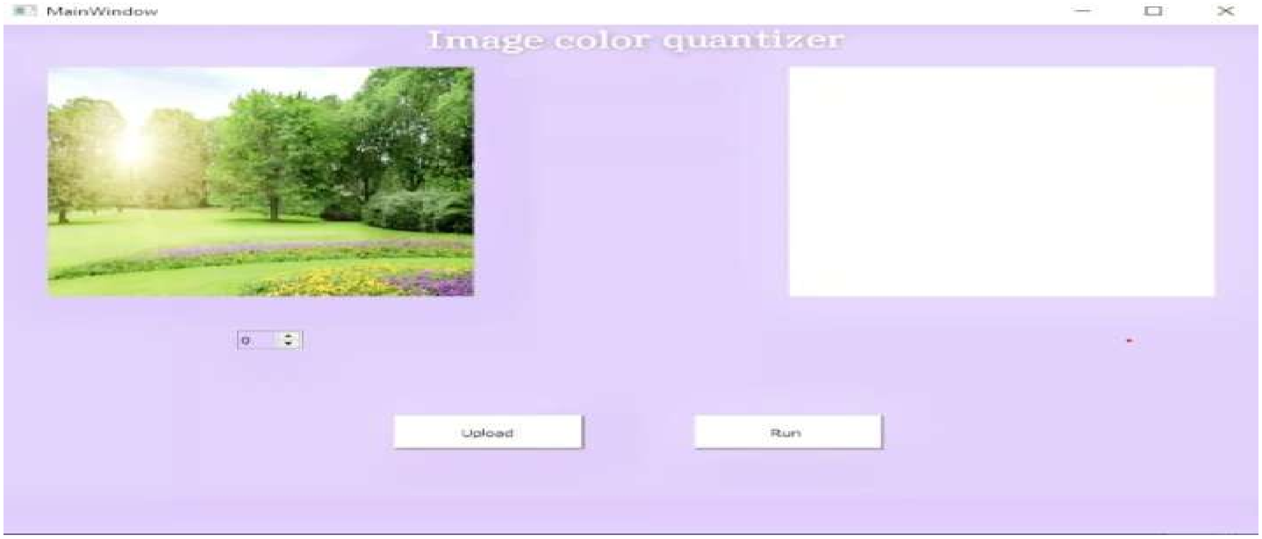


Figure 3.5: color image quantization g.u.i

3.5 Conclusion

Color quantization (CQ) is an important process in many computer graphics and image processing and analysis applications. Clustering algorithms are widely used to solve this problem. However, despite the popularity of k-Means as a general clustering algorithm, due to its high computational requirements and sensitivity to initialization, k-Means has not received much attention in the CQ literature

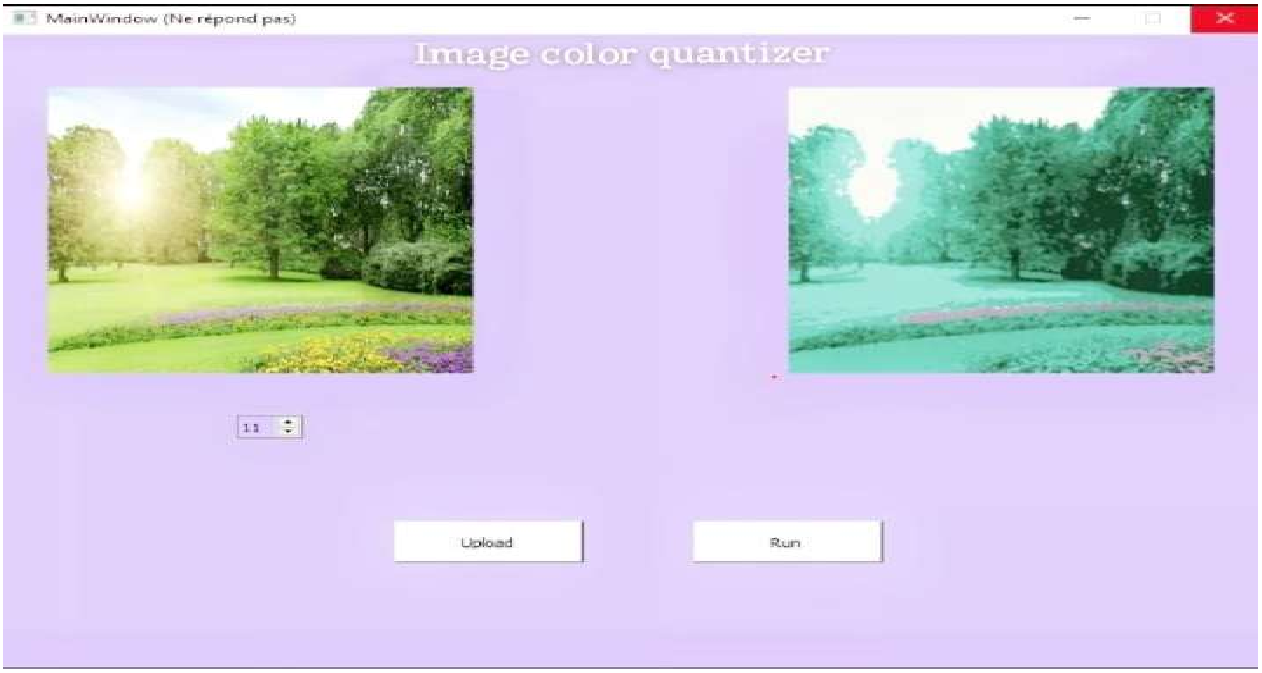


Figure 3.6: result of our application

General Conclusion

Machine learning aims at making machines capable of making decisions in many areas such as making predictions, recognizing faces and handwriting and many other tasks without being explicitly programmed. These tasks are possible by applying machine learning algorithms which it can be divided into three essential categories: Supervised, unsupervised and reinforcement learning algorithms. One of the most well-known clustering algorithms in machine learning is K-means. It is an unsupervised algorithm which offers a solution to the clustering problem. In short, it aims at forming k classes containing each of them homogeneous data points by following well-defined steps. In this manuscript, we have presented our application of K-means algorithm to quantize images in order to reduce the number of colors needed to represent it. To do that, we have implemented a python-based tool and we have showed how can image quantization produce less-color images according to the initial image

Bibliography

[1] Improved dithering methods for color quantized images. Watanabe, takashi. *Systems and computers in Japan*.

[2] A fast algorithm for color image quantization using only 256 colors. Watanabe, takashi.

[3] Color image quantization for frame buffer display. Heckbert, paul. *ACM Siggraph Computer Graphics*.

[4] KhansaaDheyaa Ismael and Stanciu Irina. Face recognition using viola-jones depending on python. *Indonesian Journal of Electrical Engineering and Computer Science*, 2020.

[5] EBM Bashier M Mohammed, MB khan. *Machine Learning : Algorithms and Applications*. International Standard Book Number, 2017.

[6] K means clustering in spatial data mining using weka interface. Sharma, r., alam, m. a., rani, a. *International Journal of Computer Applications*.

[7] Arthur Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. IBM Jornal of Research and Development, 1959.

[8] Csaba Szepesvári. *Machine Learning*. Morgan Claypool, 2009.

[9] Anthony J. Rosellini Tammy Jiang, Jaimie L. Gradus. *Supervised Machine Learning: A Brief Primer*. Behavoir Therapy, 2020.

[10] Joshi. Ameet V. *Machine Learning and Artificial Intelligence*. Springer Nature Switzerland, 2020.

[11] SamerSabri Vishal Maini. *Machine Learning for Humans*. Sachin Maini, 2017.