IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 4 (May. – June. 2024), PP 14-21
www.iosrjournals.org

Image Classification For Data Science Projects With The Help Of AI

Gyan Prabhat[1], Dr. Ashoke Kumar Mahato[2]

[1](Research Scholar, Department Of Mathematics & Master Of Computer Application, Dr. Shyama Prasad Mukherjee University, India)

[2](Associate Professor, Department Of Mathematics & Master Of Computer Application, Dr. Shyama Prasad Mukherjee University, India)

Abstract:

Image classification is termed as supervised learning problem: in which we define a set of target classes (objects to identify in images) and then train a model to recognize them using labelled example photos.

Data Science deals with collecting large amounts of data and then analyse user behaviour. The information is used to draw conclusions, make plans, implement policies, and make better, data-driven decisions.

There are several steps involved in any data science project and one of them is Data Preparation. One of the core challenges of Data Preparation is data cleaning. Research says that 80% of time and effort in Data Science projects goes to data preparation and cleaning.

AI uses machine learning technology for image recognition. AI learns by reading and learning from large amounts of image data, and the accuracy of image recognition is improved by learning from continuously stored image data.

In this paper, we investigate various types of classification algorithms suited for particular set of images. And then with the help of AI image recognition, best suited classification algorithm will be picked and applied on that image data set. This may help in minimize time and effort in data cleaning by classifying the images where data set is not classified. Any data science project can leverage the concept and build on top of this by customizing as adding or removing different classes based on the specific project requirements.

***Key Word**: Artificial Intelligence (AI); EDA; K-Nearest Neighbours; Decision Tree; SVM*

Date Of Submission: 07-06-2024 Date Of Acceptance: 17-06-2024

I. Introduction

Assigning a label or class to an input image is known as Image classification. This is part of supervised learning problem, in which a model is trained on a labelled dataset of images and their corresponding class labels. It is then used to predict the class label of new, un-seen images.

Data science studies data and how to extract meaning from it. The process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making is termed as Data Analysis. This information is then used to draw conclusions, make plans, implement policies, and make better, data-driven decisions.

The last decade has witnessed digitization of large collection of documents including handwritten manuscripts and printed documents. Recently there is an exponential increase in publicly available image databases and personal collections of pictures due to rapid growth in camera-based devices. It is estimated that over 380 billion photos were captured in the past 12 months which is 10% of all the photos ever taken by humanity.

In this context classification and extraction of information from images became an important area of research. Among all the information contained in an image text is of utmost importance, as the text carries semantic information and can provide valuable cues about the content of the image.

Basically, there are seven Fundamental Steps to Complete any Data Analytics Project [2]:

- Step 1: Understand the Business
- Step 2: Get Your Data
- Step 3: Explore and Clean Your Data
- Step 4: Enrich Your Dataset
- Step 5: Build Helpful Visualizations
- Step 6: Get Predictive

- Step 7: Iterate

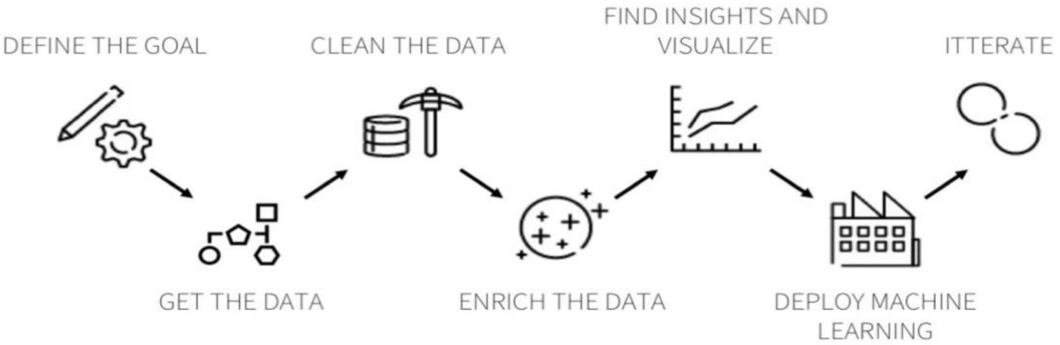

Figure 1: Steps in Data Science Project

Step 3: Explore and Clean Your Data is the dreaded data preparation process which typically takes up to 80% of the time and effort dedicated to any data science project. This research may help in minimize time and effort in data cleaning by classifying the images coming from variety of sources.

Investigating the data basically boils down to gaining a familiarity with the data of a given project. Exploratory data analysis, or EDA[9], is a particular approach to this gaining of familiarity, generally focusing on summarizing a dataset statistically and visually[8][10][11].

Image processing is the analysis and manipulation of a digitized image, usually to improve its quality. Artificial intelligence processes an image, improving the quality of an image based on the algorithm's "experience" or depth of knowledge by leveraging machine learning.

There are different types of images like medical images, satellite images, document images etc. There are various classification algorithms does exist and each is suitable for certain types of images and every algorithm have its own limitations. The idea is to investigates these classification algorithms and then by using the AI image investigation techniques choose a best suited classification algorithm on a given data set to classify based on the project requirements.

II. Out Of Box Comparison

Enterprise resource planning (ERP): It refers to a generic software system that helps run entire enterprise and supports automation. It is made of different modules: finance, supply chain, human resources, manufacturing, procurement, services, and more. It integrates different modules or business applications which talks to each other and share a common a database. ERP is a generic software solution works for all the core business processes needed to run an Enterprise: manufacturing, order management, finance, HR, supply chain, inventory, procurement, and others.

ERP helps to efficiently manage all these different modules in an integrated system. It is a generic software solution to run an enterprise.

ERP suits are critical for managing enterprises of all sizes and in all industries. Few of the ERP vendors are:

- Oracle ERP Cloud
- SAP S/4HANA
- Oracle NetSuite
- SAP ERP
- Sage Intacct
- Oracle J.D. Edwards & Company
- Workday, Inc.

When an Enterprise wish to implement ERP for better control in their business first, they need to buy the complete ERP suite. This ERP suit is generic and not designed specifically to its business needs. Hence requires an implementation and customization to fit in the business. Generic Software needs an implementation which is the process of installing, configuring, testing, and deploying a software solution to meet the needs specific to your enterprise or end-users.

And here comes the comparison. Idea is similar to build a generic approach which is suitable for any data science project where data is not classified. Then it can be customized to add/remove different classes to best suite the specific data science project needs. Hence any data science project and leverage the concept and build on top of it.

III. Classification Of Images

At its very core, from a predefined set of categories, assigning a label to an image, is termed as Image classification. Basically, this means that the task is to analyze an input image and return a label that categorizes the image. This label is always from a predefined set of possible categories.

- Medical Image
  - MRI
  - X-RAY Films
  - Ultrasound Images
- Satellite Images
  - Land
    - Coastal Lands
    - Designated Areas
    - Forest Lands
    - Grasslands
    - Deserts
    - Urban Spaces
  - Water
    - Lake
    - River
    - Sea
  - Hill
    - Fold Mountains (Folded Mountains)
    - Fault-block Mountains (Block Mountains)
    - Dome Mountains
    - Volcanic Mountains
    - Plateau Mountains
- Document Analysis
  - Handwritten Document
  - Printed Document
  - Confidential Document
  - Non-Confidential Document

Specific classification can be done based on need. For Example:

- Advisement
- Receipts
- Invoices
- Purchase Order
- Sales Order

IV. Challenges In Image Classification

There are many challenges in image classification and few of them are:

1. Intra-Class Variation
2. View-Point Variation
3. Scale Variation
4. Occlusion
5. Illumination
6. Background Clutter

1. **Intra-Class Variation:**

The variation between the images of same class is called Intra-Class Variation. Chairs of multiple types in dataset can be example of intra-class variation. A chair can be an office chair, dining table chair, comfy chair, deco chairs etc.

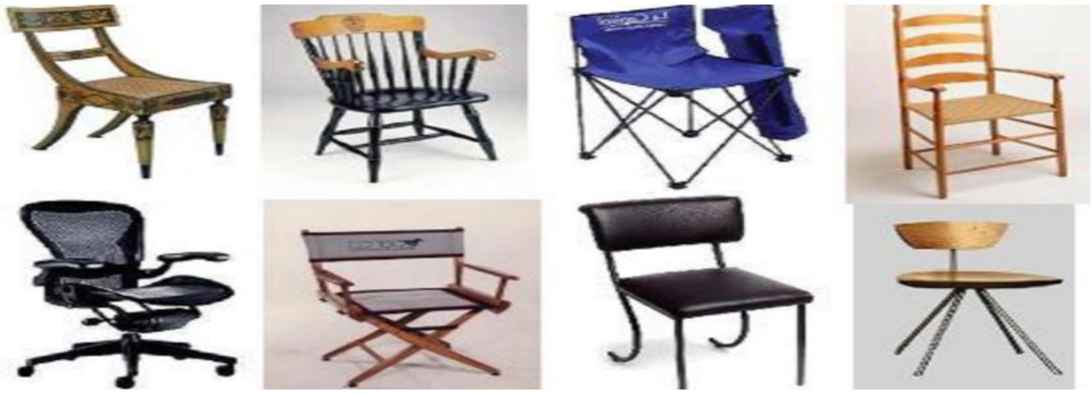

Figure 2: Intra Class Variation

2. View-Point Variation

An object can be oriented/rotated in multiple dimensions with respect to how the object is photographed and captured in image. This is termed as viewpoint variation. For example: No matter the angle in which we capture the image of chair, it's still a chair.

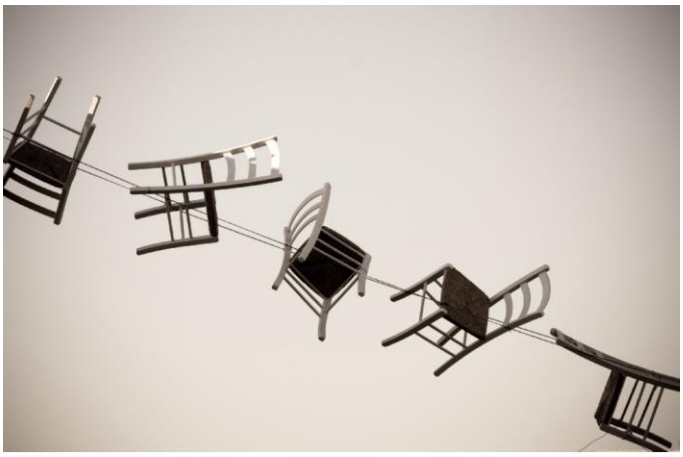

Figure 3: Viewpoint Variation

2. Scale Variation

If we have images of the same object with multiple size, then we may face scale variation challenges in image classification. This is very common in image classification.

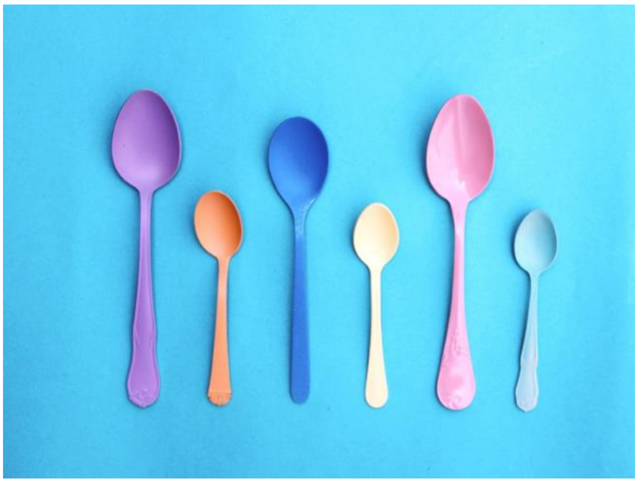

Figure 4: Scale Variation

4. Occlusion

There are a lot of objects which we want to classify in image that cannot be viewed completely. There a large part is hidden behind other objects.

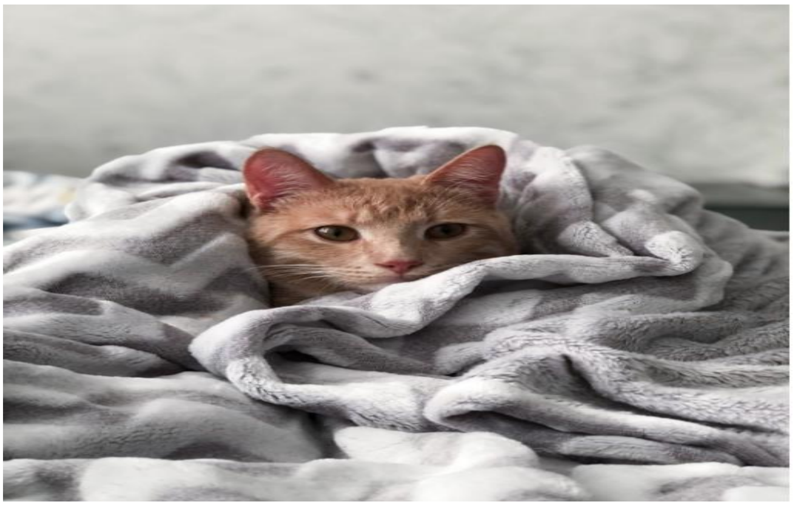

Figure 5: Viewpoint Variation

5. Illumination

Two same images can have variation in intensity level of pixels. Image classification system should be able to handle the variation in illumination. So when we give any picture of the same object with different brightness levels (Illumination) to our images classification system, the system should be able to assign them the same label.

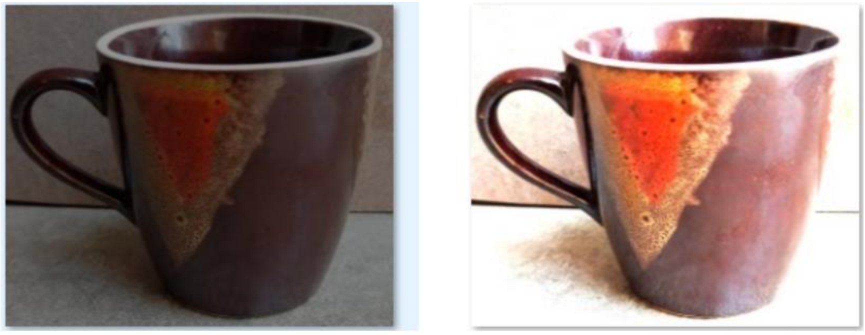

Figure 6: Illumination

6. Background Clutter

If for an observer, it is very tough to find the particular object as there are a lot of objects in the image then it is called Background Clutter. The observer is only interested in one particular object in the image; however, due to all the “noise”, it’s difficult to pick out particular object. These types of images are very “noisy” in nature.

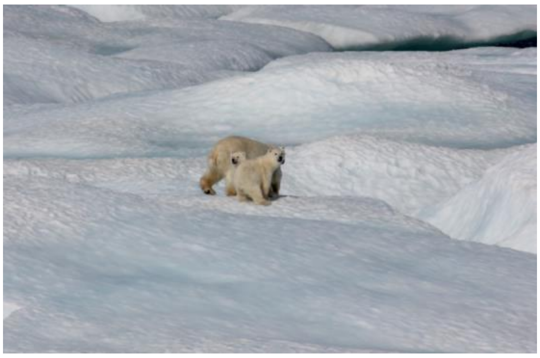

Figure 7: Background Clutter

V. Types Of Classification Algorithms

Data can be broadly divided as structured or unstructured. Classification can be performed on both types of data. It is a technique in which we categorize data into a given number of classes. To identify the category/class to which a new data will fall under is the main goal of a classification problem.

Machine learning – classification: Few of the terminology encountered:

- **Classifier:** Mapping the input data to a specific category is termed as Classifier algorithm.
- **Feature:** An individual measurable property of a phenomenon being observed is called Feature.
- **Classification model:** It tries to draw some conclusion from the input values given for training. The class labels/categories for the new data will be predicted.
- **Binary Classification:** It has two possible outcomes. Example: Gender classification (Male / Female)
- **Multi-class classification:** It has more than two classes. In this type of classification each sample is assigned to one and only one target label. Example: An animal can be tiger or cat but not both at the same time.
- **Multi-label classification:** In this each sample is mapped to a set of target labels which can be more than one class. Example: An article can be about a person, sports, and location at the same time.

The following are the steps involved in building a classification model:

- Classifier to be used needs to Initialize.
- Train the classifier: All classifiers in scikit-learn uses a best_fit(X,Y) method to fit the model(training) for the given train data X and train label Y.
- Predict the target: The predict(X) returns the predicted label Y on a given an unlabelled observation X.
- Classifier model will be evaluated.

Types of Classification Algorithms

- Logistic Regression
- Naïve Bayes
- Decision Tree
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Support Vector Machine
- Random Forest

Basic idea is to leverage above existing classification methods and then target is to define a new generic classification method that can help in classify images coming from heterogenous sources.

Logistic Regression:

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical. [1]

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumour is malignant (1) or not (0)

Naïve Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.[3][4]

Decision Tree:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. Based on comparison, we follow the branch corresponding to that value and jump to the next node. [5]

Stochastic Gradient Descent:

Stochastic gradient descent is a very popular and common algorithm used in various Machine Learning algorithms, most importantly forms the basis of Neural Networks.

Gradient descent is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function.

This algorithm is useful in cases where the optimal points cannot be found by equating the slope of the function to 0. [6]

K-Nearest Neighbours:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. [7]

Support Vector Machine:

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).[12]

Random Forest:

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. [13]

VI. Conclusion

Making a generic classification model is an idea and any data science project can be built on top of this. This concept be very useful for any data science project for which data is not classified. This generic model will classify data set and actual project can focus on the specific data that they want to use in their project. This tool can be scalable and customized based on the needs. We should add or remove different classification of data set based on the specific project requirements.

References

[1]. Juan Mario Haut1, Mercedes Eugenia Paoletti1, Abel Paz-Gallardo2,Javier Plaza1 And Antonio Plaza1, "Cloud Implementation Of Logistic Regression For Hyperspectral Image Classification", 17th International Conference On Computational And Mathematical Methods In Science And Engineering, CMMSE 2017 4–8 July, 2017

[2]. Https://Blog.Dataiku.Com/2019/07/04/Fundamental-Steps-Data-Project-Success

[3]. Aritz Pe´Rez , Pedro Larran˜aga, In˜aki Inza, " Supervised Classification With Conditional Gaussian Networks: Increasing The Structure Complexity From Naive Bayes", International Journal Of Approximate Reasoning 43 (2006) 1–25.

[4]. Sancho Mccann And David G. Lowe, "Local Naive Bayes Nearest Neighbor For Image Classification", Technical Report TR-2011-11 University Of British Columbia, December 2, 2011.

[5]. Chun-Chieh Yang, Shiv O. Prasher, Peter Enright,Chandra Madramootoo, Magdalena Burgess,Pradeep K. Goel, Ian Callum,"Application Of Decision Tree Technology For Image Classification Using Remote Sensing Data", Agricultural Systems 76 (2003) 1101–1117, Accepted 15 May 2002.

[6]. Lin Li1,Yue Wu1 And Mao Ye, "Multi-Class Image Classification Based On Fast Stochastic Gradient Boosting", Informatica 38 (2014) 145–153, December 3, 2013

[7]. L. Ma, M. M. Crawford And J. Tian, "Local Manifold Learning-Based K -Nearest-Neighbor For Hyperspectral Image Classification," In IEEE Transactions On Geoscience And Remote Sensing, Vol. 48, No. 11, Pp. 4099-4109, Nov. 2010, Doi: 10.1109/TGRS.2010.2055876.

[8]. B. Chitradevi; P.Srimathi | International Journal Of Innovative Research In Computer And Communication Engineering | Nov 11, 2014

[9]. M Manoj Krishna, M Neelima, Harshali Mane, Venu Gopala Rao Matcha |International Journal Of Engineering & Technology 7(2.7):614 |March 2018

[10]. V. Frinken A. Fischer, A. Keller And H. Bunke. Hmm-Based Word Spotting In Handwritten Documents Using Subword Mod- Els. In ICPR, 2010.

[11]. Huawu Deng ; D.A. Clausi |IEEE Transactions On Pattern Analysis And Machine Intelligence (Volume: 26 , Issue: 7 , July 2004)

[12]. Tzotsos, A., Argialas, D. (2008). Support Vector Machine Classification For Object-Based Image Analysis. In: Blaschke, T., Lang, S., Hay, G.J. (Eds) Object-Based Image Analysis. Lecture Notes In Geoinformation And Cartography. Springer, Berlin, Heidelberg. Https://Doi.Org/10.1007/978-3-540-77058-9_36

[13]. W. Man, Y. Ji And Z. Zhang, "Image Classification Based On Improved Random Forest Algorithm," 2018 IEEE 3rd International Conference On Cloud Computing And Big Data Analysis (ICCCBDA), Chengdu, China, 2018, Pp. 346-350, Doi: 10.1109/ICCCBDA.2018.8386540.