

IMAGE SCRAPING AND CLASSIFICATION PROJECT

Submitted by:

DEBANTI ROY

**ACKNOWLEDGMENT**

I, Debanti Roy, would like to convey my sincere gratitude to DataTrained Academy and Flip Robo Technologies for giving me this opportunity to do this project. I would like to thank all mentors and SME’s for extending their support all through the process which helped me complete this project.

**INTRODUCTION**

* Business Problem Framing

The idea behind this project is to build a deep learning-based Image Classification model on images that will be scraped from e-commerce portal. This is done to make the model more and more robust.

* Conceptual Background of the Domain Problem

Images are one of the major sources of data in the field of data science and AI. This field is making appropriate use of information that can be gathered through images by examining its features and details. We are trying to understand how an end to end project is developed in this field.

* Review of Literature

Image classification is an important step in the object detection and image analysis. The output of the image classification step can be the final output or the intermediate output. A lot of image classification techniques have been proposed till date. Various studies have been conducted in order to conclude about the best satellite image classification technique. It is hard to decide any one technique as the best technique among all, because the results and its accuracy depend on a number of factors.

Over the last few decades, there is a constant modification in the conventional methods as well as invention of new image classification techniques in order to get maximum accurate results. Each of the classification technique has its own advantages and disadvantages. The research now concentrates on combining the desired features of these techniques in order to increase the efficiency.

* Motivation for the Problem Undertaken

The objective of image classification is to identify and portray, as a unique gray level (or color), the features occurring in an image in terms of the object or type of land cover these features actually represent on the ground. Image classification is perhaps the most important part of digital image analysis.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

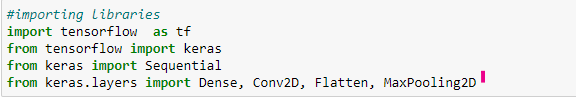
* Data Sources and their formats

We have scraped images of sarees (women), jeans (men) and trousers (men) from an ecommerce site (Amazon) for training the data. For each category we have collected 250 images and have saved them as images and the record was stored in csv format.

We have also collected images separately for testing from websites like Flipkart and Amazon and saved them as images. The records are saved in csv format. Here each category consists of 250 images.

* Data Preprocessing Done
* **Importing the necessary libraries and packages**

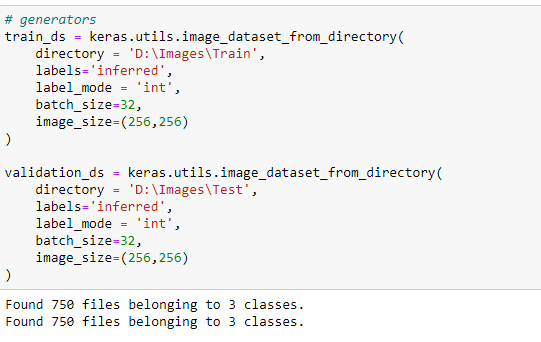
The first step to process any data is to first import the necessary packages required for the same. We have imported the necessary libraries and packages for preprocessing and testing as follows:



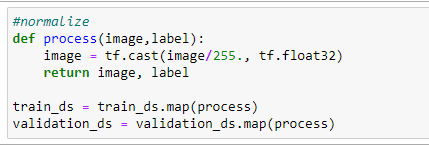
Henceforth, we have to import the required dataset. Our data is present in a zip file. Thus, we have used the following methods to import:

****

As a part of preprocessing the data we made sure that all the images both in training and test dataset is of the same shape and size. In order to ensure that we have used generators to make all the images of equal shape and size.



We applied normalization to change the range of pixel intensity values to make the image more familiar.



* Data Inputs- Logic- Output Relationships

Both the training and test datasets are saved as images and we are using the train dataset to feed out machine and understand how each category of images looks like.

The test dataset is further used to validate the performance of the model as to how the model interprets any new data.

* Hardware and Software Requirements and Tools Used

Hardware required:

* 1. Processor: core i5 or above
  2. RAM: 8 GB or above
  3. ROM/SSD: 250 GB or above

Software required:

* 1. Anaconda 3- language used Python 3
  2. Microsoft Excel Libraries: The important libraries that I have used for this project are below:

*import numpy as np*

It is defined as a Python package used for performing various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

*import pandas as pd*

Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. The data-frame term is coming from Pandas only.

*import matplotlib.pyplot*

Matplotlib acts as the backbone of data visualization through Python.

* **Matplotlib**: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statical interferences and plotting 2D graphs of arrays.

**Model/s Development and Evaluation**

* Testing of Identified Approaches (Algorithms)

We have applied Conv2D, MaxPooling2D and Dense for building the model. Our model performance was \_\_\_\_\_\_\_\_\_\_\_\_\_at the end.

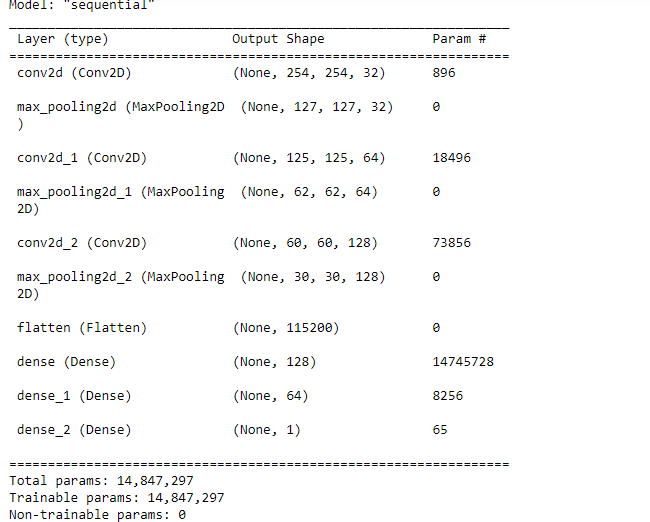
Therefore, to improve the performance we applied BatchNormalization and Dropout. The performance improved to \_\_\_\_\_\_\_\_\_\_.

* Run and Evaluate selected models

We have built a CNN model to train and then evaluate the performance model.



We checked the model structure with the help of model summary.



* Visualizations

We have visualized the model performance using a

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

* Interpretation of the Results

|  |  |
| --- | --- |
| **CNN Model Performance** | |
| Before applying BatchNormalization & Dropout |  |
| After applying BatchNormalization & Dropout |  |

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

This project helps us understand that CNN has a greater scope in Machine Learning and can offer more accurate results with more and more data. We can see the performance of our model has increased from \_\_\_\_\_\_\_\_\_\_\_\_\_\_ to \_\_\_\_\_\_\_\_\_\_\_\_, which is a good improvement.

* Limitations of this work and Scope for Future Work

As we are working with a small dataset results might not be much accurate. However, CNN models are more effective with larger datasets.