

Image Scraping and Classification Project

Amazon

Submitted by:

Narendran Suresh

**ACKNOWLEDGMENT**

I would like to take the opportunity to thank the organisation ‘FlipRobo Technologies”, “Datatrained “and my mentor Ms. Astha Mishra for their immense support and guidance for making this feat achievable. I would also like to thank my dad Mr. Suresh Kumar S who is a statistics subject expert for guiding me along the right path and techniques to be used.

FlipRobo Technologies who are specialised in making ML / AI models provided me the Data set as a part of my internship.

References were made to several articles among Medium, KdNuggets, towardsdatascience, realpython, machinelearningmastery, python and sklearn documentation for the successful completion of the project.

**INTRODUCTION**

* Business Problem Framing

To scrape images of apparel from the web ( Amazon.com) for three categories (Saree, Mens Jeans and Mens Trousers), and to develop a deep learning algorithm to classify between the objects.

* Conceptual Background of the Domain Problem

TO be able to classify images into different categories either if the data is scraped for a purpose or to do the same on the existing image database in personal systems could lead to a efficient filing system. Image classification is also used for customer ease and for companies to take informed decisions.

* Review of Literature

Extensive reseach was done on CNN and Residual Network in order to find the solution. Web scraping domain knowledge and research was done to enhance the techniques used for data retrieval and image format conversion.

* Motivation for the Problem Undertaken

A lot of data is available in the form of images / videos and the capability of machines to predict / classify objects based on the image properties is a domain where deep learning and neural networks can play a major role in saving time and efforts for humans.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The use of residual networks was critical for arriving at an accurate solution for the problem in hand.

Image conversion to arrays and the normalization techniques were critical for arriving at the solution.

* Data Sources and their formats

No data was provided for the scope of the project. Instead, it was scraped from amazon using web scraping tools selenium and beautifulsoup. The collected images were saved to a destination folder, extensions were added and split for the purpose of classification

* Data Preprocessing Done

Data was split in the ratio of 60:20:20 against training, validation and testing.

Residual Network (ResNet50) was used as a base model and the actual model was built on top of the residual network.

* Data Inputs- Logic- Output Relationships

The data was pretty trivial and the classification prediction task was attained by the use of Convolutional neural networks. The challenge was that trousers and jeans could be easily mistaken for each other, but the use of ResNet helped in solving the problem

* State the set of assumptions (if any) related to the problem under consideration

N/A

* Hardware and Software Requirements and Tools Used
* Hardware:
* Inter Core (i7) – 5500U, clock speed at 2.40GHz
* RAM – 12.0 GB
* Software:
* Jupyter Notebook (Anaconda 3) – Python 3.7.6
* Microsoft Excel

Libraries & Packages used – Pandas, numpy, sklearn, matplotlib, seaborn, sklearn, scipy, imblearn, tensorflow, keras, Image Data Generator, Maxpool2D

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Scraping the images, the initial task was pretty straight forward. The use of for loops to loop between the desired objects and saving them to a root folder and changing their extensions were done.

Upon development, the model using CNN alone with optimizer ‘adam’ and ‘RMSprop’ did not show much improvement to the model with the loss set at categorical cross entropy at first.

The use of Residual Network helped solve the problem and reach a good level of accuracy.

* Testing of Identified Approaches (Algorithms)

**CNN Optimizers:**

Adam

RMSProp

**ResNet50**

**Base model was set as resnet with parameters as below:**

base\_model=ResNet50(include\_top=False)

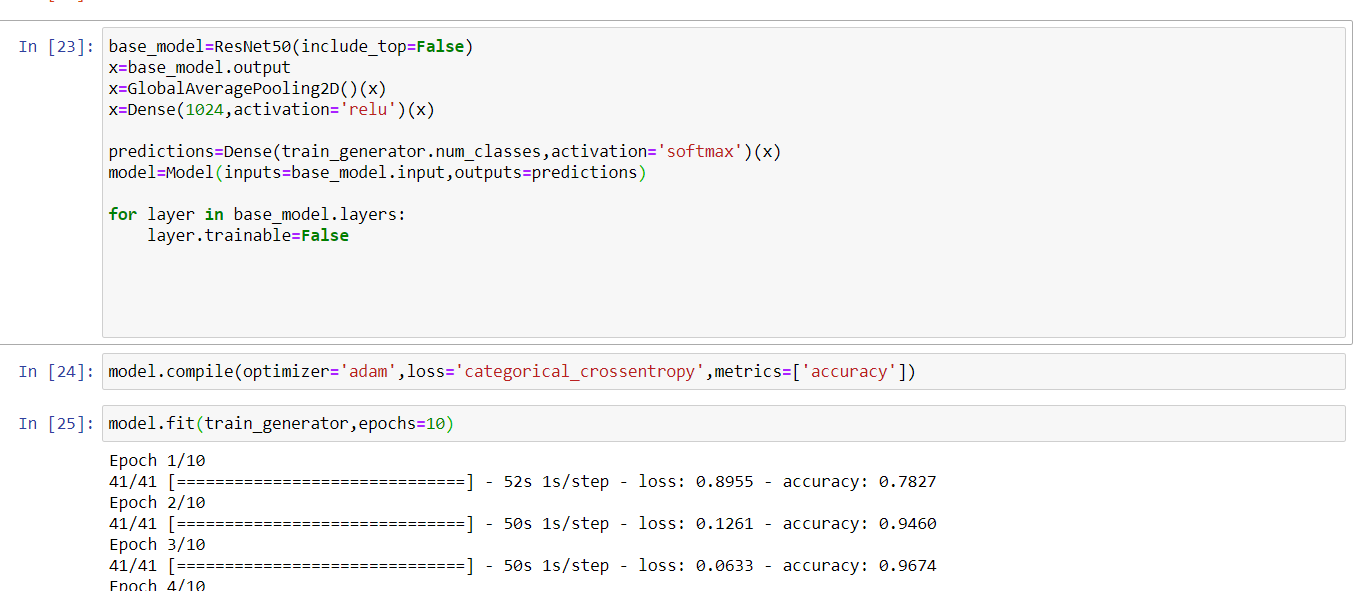
x=base\_model.output

x=GlobalAveragePooling2D()(x)

x=Dense(1024,activation='relu')(x)

* Run and Evaluate selected models

CNN and Resnet were used in combination to arrive at the solution.

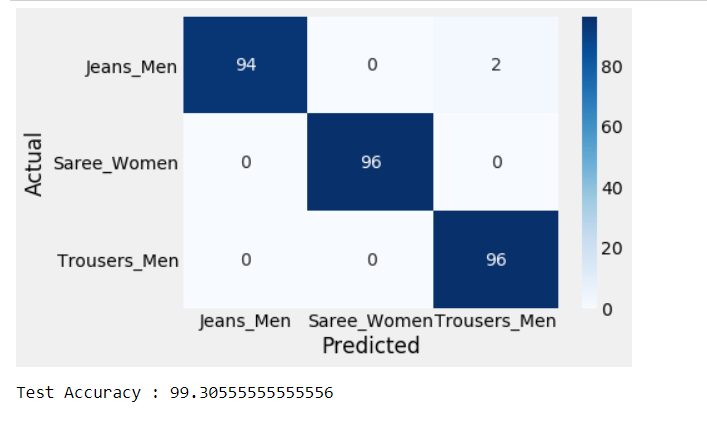


* Key Metrics for success in solving problem under consideration

Accuracy was the key metric for the solution to the problem. Upon use of ResNet, the loss was found to be dropping drastically.

* Visualizations

The confusion matrix of the model shows that the model in hand is a very dependable one and can be deployed for solution to the problem.



* Interpretation of the Results

The model does a very good job in classifying the dress categories and can be used for predictions. Addition of more categories would mean that the model can be used for a wide range of applications.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The images scraped were saved as files and not jpeg or jpg which had to be processed again.

* Learning Outcomes of the Study in respect of Data Science

The use of ResNet50 and understanding about the CNN networks like LeNet, AlexNet and VGG.

* Limitations of this work and Scope for Future Work

The data for some categories were limited and hence model may not perform as well if more data is added.

Addition of more data would mean the number of batches will have to be increased and hence the computational time might increase drastically for a decent number of epochs.