Image Scraping and Classification Project

Submitted by:

Sumit Santra

ACKNOWLEDGMENT

I would like to express my special thanks and gratitude to my SME **Shubham Yadav**, Who gave me the golden opportunity to do this wonderful project **Image Scraping and Classification Project**.

I would like to thank them to give me Such a wonderful project to learn and test my knowledge on and I learnt a lot of things from this project.

INTRODUCTION

- Business Problem Framing
  - ➔ 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 give you an exposure of how an end to end project is developed in this field.
  - ➔ 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
  - ➔ Here we need to understand how a Convolution Neural Network Works.
  - ➔ We need to have a knowledge where to apply which cost function, Optimizer and Which model should be used to create the model more effective in classifying the images.
  - ➔ Here images takes the classes as the folder like if there are 3 folders i.e there are 3 classes and we have to classify 3 classes based on the image data we have
  - ➔ In Convolution Neural Network we need more data as it need to extract features from that image so that it can classify the images based on the category.

- Review of Literature
  - ➔ First I started research about CNN and how it works then I read about about it Collects features from the images and how it classifies the images.
  - ➔ How the filter slides on the images and how it takes important features from the images.

- Motivation for the Problem Undertaken
  - ➔ My motivation to undertake this Project is my mentor Shubham Yadav Sir & to get the knowledge about the Web Scraping and Image

Classification So that it will be easy for me to undertake this kind of project easily as I know the Concepts behind Convolution Neural Network and how to pre process the data and how to classify the images.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

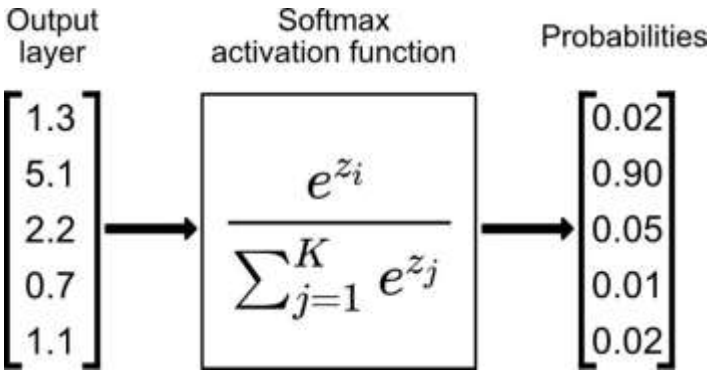
➔ Activation Function:

1. Relu Activation Function:

$$ReLU$$

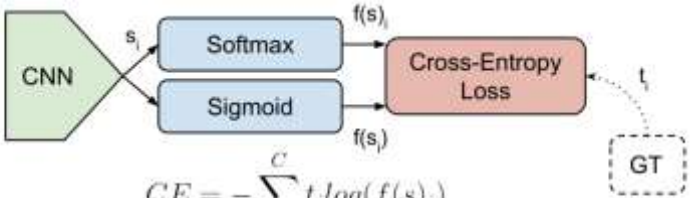
$$f(x) = max\ (0, x)$$

2. Softmax Activation Function:



➔ Cost Function:

1. Categorical Cross Entropy:



$$CE = -\sum_{i}^{C} t_i log(f(s)_i)$$

$$CE = -\sum_{i=1}^{C'=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1 - t_1) log(1 - f(s_1))$$

➔ Optimizer:

1. Adam Optimizer:

$$m_t = (1-\beta_1)\sum_{i=0}^{t}\beta_1^{t-i} g_i$$

➔ This are the Mathematical things that we have used in the project.

- Data Sources and their formats

➔ Data Source is a Image they are saved in .png format

➔ I Scraped the data using Selenium and Python from Flipkart and Store them in different different directories so that the model can classify them based on the folders.

➔ I scraped three categories i.e Saree, Jeans Men, Trouser Men.

➔ Scraped 1000 images for each Category and saved them to there respective folders.

➔ From this 1000 images some images I transferred to test set to use my model to classify them on there own.

➔ The Data looks like this:

1. Saree:

2. Jeans Men:

3. Trouser Men:

➔ This is how the data looks we have this kind of data in the dataset.

- Data Preprocessing Done
  - ➔ At first I did data augmentation using Imagedatagenerator.
  - ➔ After that I divided the data by 255 so that the data gets normalized and it does not take more space on our ram. By doing this the data has all the pixels in between 0-1.
  - ➔ By adding more data using data agumentation I tilt the data, zoom in the image so that the data can increase and our Model can learn more.

- Data Inputs- Logic- Output Relationships
  - ➔ The relation between the Train and test data are same as they both are images we are first training the data and then we are testing the data on test set so that our model can classify the images on its own.
  - ➔ As it is a images data our model learns some features from the images and if we put wrong images in wrong folders then our model can learn incorrect features and it may give us a wrong answers.

- State the set of assumptions (if any) related to the problem under consideration
  - ➔ I first thought that it will be easy to scrape images from the e-commerce site by because of load on flipkart it was giving me errors so it took time for that to solve.
  - ➔ Then I scraped 500 images and I was not getting that good accuracy for my model so after that I scraped all the pages of the particular model and after scraping all the images my model gave 99.24% accuracy.
  - ➔ So in this step I only used try and error technique to get a good accuracy.

- Hardware and Software Requirements and Tools Used
  - ➔ **Hardware Used:**
    1. CPU: RYZEN 9 5900HX.
    2. GPU: RTX 3060.
    3. RAM: 16GB.
    4. Storage: 1TB SSD.

  - ➔ **Software Used:**
    1. OS: Windows 10.
    2. IDE: Jupyter Notebook.

  - ➔ **Python Libraries Used:**
    1. Numpy

2. Matplotlib
3. Seaborn
4. Os
5. Tensorflow
6. Keras
7. glob

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  - ➔ As we know it is a image dataset so when I was scraping the dataset at that time using the OS library I created separate directories for different product images if I did't create different directory then it would have been a problem as the model will only have one category so I created 3 folders and store images in the respective folders.
  - ➔ As images are Coloured images so they have 3 layers Red, Blue, Green and each layer has 0-255 pixels so if I load the data with 0-255 pixel density they my ram would not be sufficient so I divided all the images with 255 so that the pixel density of all the images come down between 0-1 and because of this our model will work fast.
  - ➔ As we know Convolution Neural Network needs more data so I used Data Augmentation Technique to increase the data so that our model could learn every features of our Images.

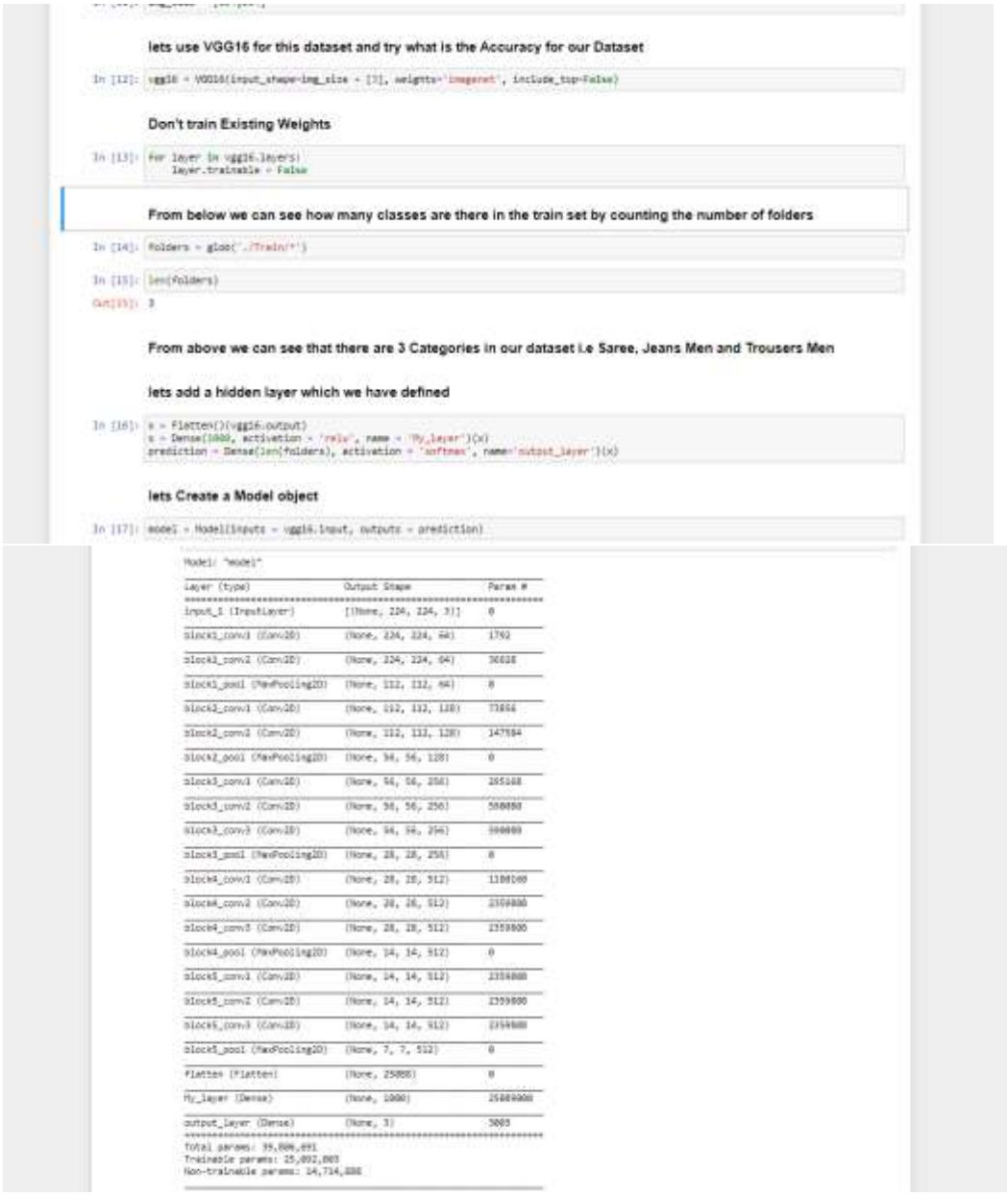
- Testing of Identified Approaches (Algorithms)

  Listing down all the algorithms used for the training and testing.

  - ➔ VGG16
  - ➔ Inceptionv3
  - ➔ Resnet50

- Run and Evaluate selected models

➔ VGG16



lets use VGG16 for this dataset and try what is the Accuracy for our Dataset

```
In [12]: vgg16 = VGG16(input_shape=img_size + [3], weights='imagenet', include_top=False)
```

Don't train Existing Weights

```
In [13]: for layer in vgg16.layers:
             layer.trainable = False
```

From below we can see how many classes are there in the train set by counting the number of folders

```
In [14]: folders = glob('./Train/*')

In [15]: len(folders)
Out[15]: 3
```

From above we can see that there are 3 Categories in our dataset i.e Saree, Jeans Men and Trousers Men

lets add a hidden layer which we have defined

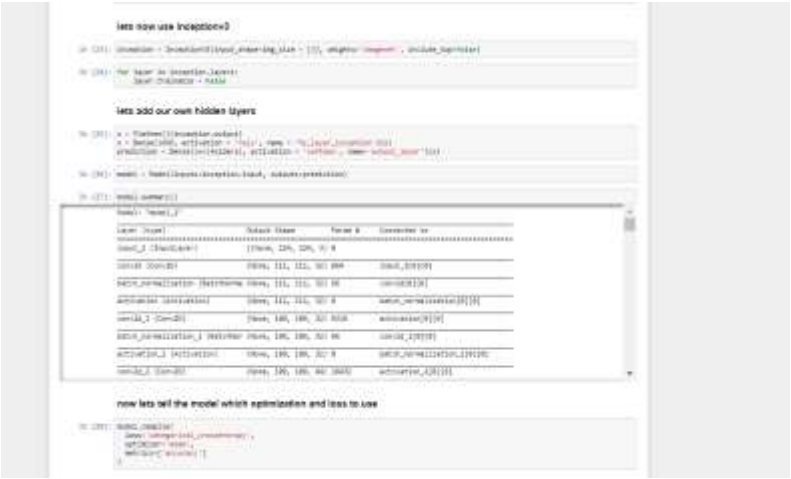
```
In [16]: x = Flatten()(vgg16.output)
         x = Dense(1000, activation = 'relu', name = 'My_layer')(x)
         prediction = Dense(len(folders), activation = 'softmax', name='output_layer')(x)
```

lets Create a Model object

```
In [17]: model = Model(inputs = vgg16.input, outputs = prediction)
```

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 224, 224, 3)]     0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 56, 56, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 56, 56, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 28, 28, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 28, 28, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 28, 28, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 28, 28, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
_________________________________________________________________
Flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
My_layer (Dense)             (None, 1000)              25089000
_________________________________________________________________
output_layer (Dense)         (None, 3)                 3003
=================================================================
Total params: 39,806,691
Trainable params: 25,092,003
Non-trainable params: 14,714,688
_________________________________________________________________
```

➔ Inception V3



➔ Resnet50

- Key Metrics for success in solving problem under consideration
  - ➔ Loss: - The Loss Function is one of the important components of Neural Networks. Loss is nothing but **a prediction error of Neural Net**. And the method to calculate the loss is called Loss Function. In simple words, the Loss is used to calculate the gradients. And gradients are used to update the weights of the Neural Net.
  - ➔ Validation Loss: - Validation loss (as mentioned in other comments means **your generalized loss**) should be same as compared to training loss if training is good. If your validation loss is lower than the training loss, it means you have not split the training data correctly

- Visualizations
- Saree:

- Jeans Men:

- Trouser Men:

- Interpretation of the Results
  - ➔ By looking at the visualization we can easily say that there are three categories and there names are Saree, Jeans and Trousers
  - ➔ We can see the images of all the 3 categories and we have a total of 1000 images for each categories

CONCLUSION

- Key Findings and Conclusions of the Study
  - ➔ From this project I understand that the Deep Neural Network or Convolution Neural network needs large data to perform well when I gave it 500 images data it was performing good but not as good as 1000 images which I passed in later point of time.
  - ➔ And we need to normalize the image data so that the pixel density is between 0-1 and we can do Convolutions faster.

- Learning Outcomes of the Study in respect of Data Science
  - ➔ From this project I came to know how to scrape images from the website easily and how can we create different folders using OS library if the folder is not present in the location.
  - ➔ I came to know that it is important to normalize the images so that we can do the Convolutions Faster.
  - ➔ I came to know how to add a GPU while using TensorFlow and Keras.

- Limitations of this work and Scope for Future Work
  - ➔ This model has some limitations as this model is only trained on 3 categories of products so it can only classify 3 categories if we pass other product images it can classify Wrongly.
  - ➔ To improve the model accuracy we can train the model on more data from different websites.
  - ➔ To teach our model more categories we need to feed our model with more categories and our model will perform good on other product as well.