PROJECT REPORT

ON

IMAGE SCRAPING AND CLASSIFICATION

Submitted by

MANISHA

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 an exposure of how an end to end project is developed in this field.

Conceptual Background of the Domain Problem:

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. This task is divided into two phases such as Data Collection and Model Building.

Review of Literature:

We need to scrape images of these 3 categories and build your data from it. That data will be provided as an input to your deep learning problem. We scrape minimum 200 images. We apply image augmentation techniques to increase the size of your data but make sure the quality of data is not compromised.

Motivation for the Problem Undertaken:

After the data collection and preparation is done, we build an image classification model that will classify between 3 categories. And also optimizers the learning rates for improving our model's performance.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

We use Statistical techniques and analytics modeling in our projects, such as:

- Convert image into matrix form which can be easily readable.
- corr(): use to calculate the relation between all the categories.
- Resize the image size to get proper quality.

Data Source and their formats:

There are 216 rows and 2 columns. ‘Categories’ column is our target variable and image data is feature variable. In target variable there are three categories which are ‘Sarees’, ‘Trousers (Men)’, ‘Jeans (Men)’.

Data Pre-processing

In pre-processing we convert image data into numeric form and the resize the image so image quality become better or equal in size for all the images.

Hardware and Software Requirements and Tools used

Hardware:

- Memory 16GB minimum
- Hard Drive SSD is preferred 500GB
- Processor intel i5 minimum
- Operating system Windows 10

Software:

- Jupyter notebook (Python)

Libraries:

pandas (used to create the data and read the data)

tensorflow (used to load the images)

numpy (used with the mathematical function)

resize (used to improve the size)

seaborn (used to create a different types of graphs)

matplotlib (used to plot the graph)

accuracy_score (used to calculate accuracy score for train and test)

classification_report (to display precision, f1 score)

confusion_matrix (form the matrix)

Model/s Development and Evaluation

Identification of possible problem:

We approach to both statistical and analytical problem

- ❖ Plot a bar graph for nominal data
- ❖ corr() used to calculate the correlation of input variable with the output variable.

Testing of Identified Approaches:

Here we work on the classification problem so the machine learning models are:

- Logistic Regression
- K Neighbors Classification
- Random Forest Classification
- Decision Tree Classification
- SVM

Run and Evaluate selected models:

- ➢ Logistic Regression

The accuracy score of logistic regression is 52. The total of true negative and false negative in the confusion matrix is 54.

- ➢ K Neighbors Classification

The accuracy score of K Neighbors classification is 74. The total of true negative and false negative in the confusion matrix is 30.

- Random Forest Classification

The accuracy score of Random Forest classification is 92. The total of true negative and false negative in the confusion matrix is 17.

- Decision Tree Classification

The accuracy score of Decision Tree classification is 87. The total of true negative and false negative in the confusion matrix is 22.

- SVM

The accuracy score of SVM is 96. The total of true negative and false negative in the confusion matrix is 11.

The SVM model gives better accuracy score. Precision score, recall and f1-score is also good in compare to the other. The total of True Negative and False Negative in the confusion matrix is also less in the same model.

Visualisation:

On visualising the data we see the data all the three categories are equal in number which means our target variable is balanced.

Interpretation of the Results:

On our analysis basis we go through various models and then we conclude better model on the basis of accuracy score. That will predict the test data on training the train data. This shows that 96% prediction is correct and only 4% is wrong.

Conclusion

Key Findings and Conclusions of the Study:

This is an image classification project, where we work on Deep Learning. We convert image data into matrix form by which data can become readable and also convert three categories. All the categories are in equal number. We resize the data also in it.

Learning Outcomes of the Study in respect of Data Science:

After learning the data and models we come to know that SVM is the better model for this project because accuracy score is maximum for it and also have least total of true negative and false negative in confusion matrix.