IMAGE SCRAPPING AND CLASSIFICATION PROJECT

Submitted by:

ANITA THAPA

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my SME Swati Ma’am and FlipRobo for giving me an opportunity to work on the Image Scrapping and Classification Project. During my research on this project internet sites like towardsdatascience, medium, Amazon.in, scikit learn and Stack Overflow helped me a lot in the 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.

- ## Review of Literature

  A bit of Research was done on the topic of CNN in Deep Learning, which helped in understanding the Image classification project.

- ## Motivation for the Problem Undertaken

  The project was assigned as an internship project for better understanding of topics like Image classification, CNN and Deep learning.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  A total of 930 Images were scrapped from e-commerce portal, Amazon.in. The clothing categories used for scraping are:

- Sarees (women)
- Trousers (men)
- Jeans (men)

- Data Sources and their formats

  The scrapped data are images of the format '.jpg' which are stored in a folder.

- Data Preprocessing Done

  Scrapped was saved in local folder, later it was uploaded in google drive for performing modelling in google colab.

- Hardware and Software Requirements and Tools Used

  Hardware:

  The system with a 16-core processor has been used, the operating system was Windows 11, Google colab has been used for performing Deep learning:

  Machine Learning Libraries:

  Pandas, Numpy, seaborn, matplotlib, keras, Dense, Input, Dropout, GlobalAveragePooling2D, Flatten, Conv2D, BatchNormalization, Activation, MaxPooling2D, Adam and accuracy_score

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  In google colab new directory named ‘dataset_amazon’ was created under which train and test directory was created in which scrapped images were filled randomly. Test folder was given 20% of the images.

- Testing of Identified Approaches (Algorithms)

  CNN model has been used which gave 97% accuracy

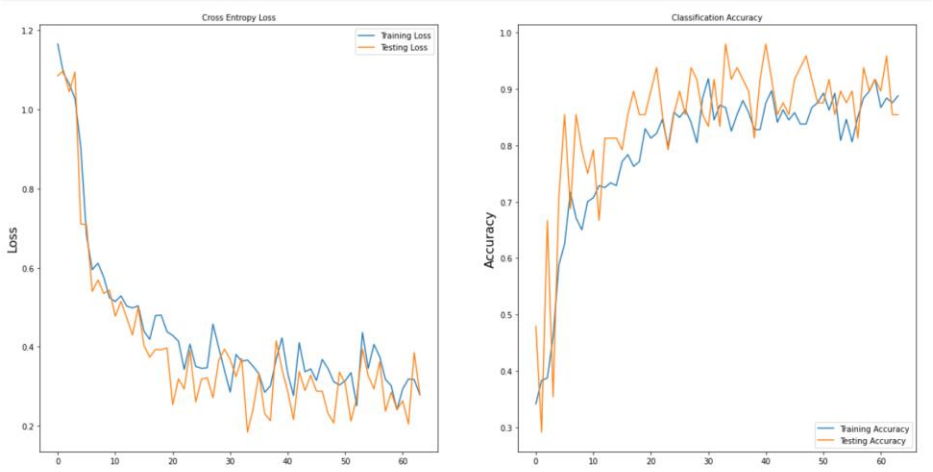
  4 convolutional layers and 2 fully connected layers were used.

- Run and Evaluate selected models

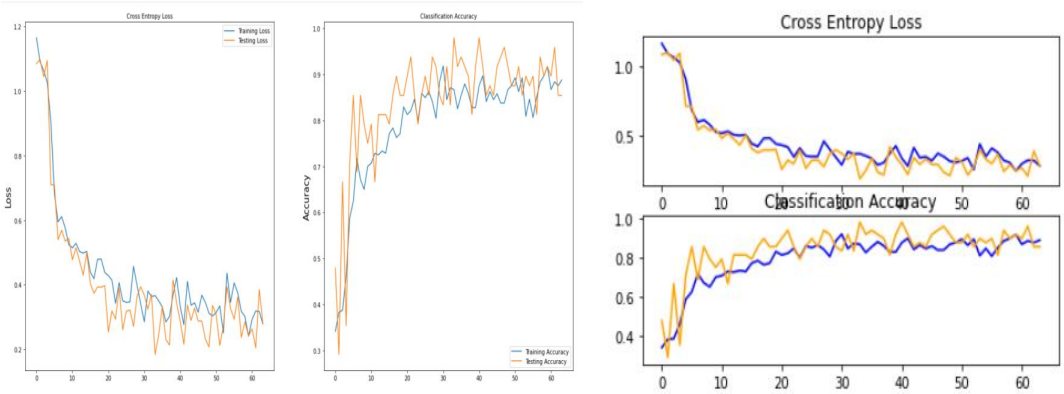
  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- Key Metrics for success in solving problem under consideration

  Accuracy and loss is used as key metrics as shown in the graph below.



- Visualizations



- Interpretation of the Results

  Model approximately predicted the correct output but, in some images, model is not able to predict correct values as it is due to small dataset.

CONCLUSION

- Key Findings and Conclusions of the Study

  Classification model was able to predict and distinguish between the 3 categories namely: trousers, jeans and saree. Model has an accuracy of 97%

- Learning Outcomes of the Study in respect of Data Science

  Learned about the CNN and deep learning topics and as per the predicted value it is inferred that more the data better will be model's performance.

- Limitations of this work and Scope for Future Work

  Use of random images confuses our model between the classes but it can be overcome by using large training dataset.