ML Project

By VARPAS

Short Report: (4-5 pages max)
a. Introduction: Outline your approach and understanding of the problem.
b. Methods: Briefly describe the tools, libraries, and techniques you employed.
c. Results: Show extracted data snippets, visualizations, and insights from the data analysis.
d. Conclusion: Highlight your findings and any challenges you encountered.

Reference-- https://nanonets.com/blog/ocr-with-tesseract/#tesseract-ocr

Objective of the project

MNIST dataset popular dataset of of handwritten digits in the field of image processing. It contains grey style images from 0 to 9.

Scope of the project

Step1- Randomly select 100 images

step2- Add noise to images means add anything i.e. like line, dots, scribble to that image (ruin the image).

step3- Install pytesseract and Tesseract OCR to identify images. -- These are open source tools. First is for better language detection and later is user for extracts text from images and documents without a text layer

step4- . Calculate the accuracy of your OCR method on these noisy images.-- here we got 98% accuracy on training dataset and 99% model accuracy on testing dataset. Check misqualified digits and mis-qualification rate

Key Learnings

1. Python libraries to handle the OCR image processing (pytesseract, tesseract)
2. Tried to use mnist dataset as per their guidelines, but couldn’t open the images. So need to find out option that without downloading how to access data.
3. Model Accuracy on training dataset and test dataset is above the threshold so we can register this model for further project development.

Thank You!