Group 146 Project Report: textOCR

Dhruv Chand, Roshaan Quyum, Arin Khandelwal
{chandd9,quyumr,khanda3}@mcmaster.ca

1 Introduction

- We are creating an application that extracts text from an image, otherwise known as Optical Character Recognition (OCR). It is a multiclass classification task using images with text on it as its input data. It can also be seen as a single label classification task where each data point will be classified into only one class representing the character that it is. We will create a command line application that takes an image with English text as input and outputs the text on the image into the terminal. It involves implementing a number of different parts of the machine learning and model training process such as image preprocessing, text recognition, and character recognition. We shall use machine learning libraries such as OpenCV, NumPy, Scikit-Learn and PyTorch to implement them. If we have the time, we would like to extend the OCR to recognize text in natural photographs (e.g. recognize text in a photo of a stop sign).

2 Dataset

- from the sheet. check the data set from. refer to the new data set.

We chiefly used two datasets: OCR-Dataset and Words MNIST.

OCR-Dataset consists of the 26 letters of the English alphabet in both lower and uppercase along with all 10 digits; across the 62 classes, there are roughly 210 000 images in total. The data uses 3475 different fonts from publicly available Google Fonts. It is the dataset we trained our model on. There are a number of preprocessing steps done to the images from this dataset. Images are converted to a specific type (`np.uint8`) to ensure consistency in the following preprocessing steps. Then, all white pixels on the edge of the letter are cropped out (the cropping is similar to a bounding box being placed around the letter/digit and then all pixels outside of it removed from the image). The image is then scaled to have dimensions of 64x64, and finally, it is turned into a tensor. After this preprocessing, the images from this dataset can be inputted to the model. This preprocessing here assumes the image contains a single character; however, for images with more than one character, extra preprocessing was needed.

The Words MNIST dataset

3 Features and Inputs

- diff features and inputs for pre processing.

4 Implementation

5 Evaluation and Progress

- Copy the description of the old model.

6 Error Analysis

- Need to figure out

Team Contributions

7 Figures and Tables