CSC730: Report for Assignment 1

Jacob James, David Mathews, Kris Jensen

January 15, 2024

1 Introduction

The first assignment of this course tasked us with analyzing a set of skewed data from the MNIST dataset. This skewed dataset contained eight classes from a total set of ten classes. Also, each class was not represented with equivalent frequency. The dataset contains 12244 records of data. Each record contains a 784-element list that represents a 28x28 image of a handwriting sample.

This report will detail the steps taken by our team to generate an anomaly score for each image and compare our results for accuracy.

We used two primary toolsets to perform the analysis. One toolset was python executed on VS code and the other was python executed on Google Colab.

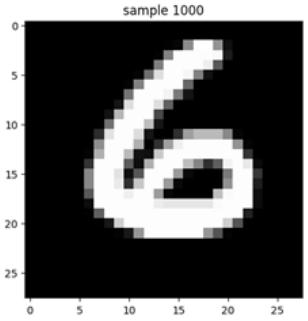

Figure 1: Handwriting Sample 1000

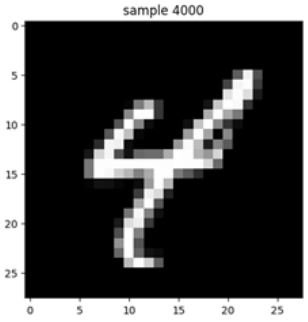

Figure 2: Handwriting Sample 4000

2 Methodology

The methodology that provided the best results for this assignment counted the the pixels that exceeded an arbitrarily chosen threshold value of 128. The maximum gray scale intensity being 255, this value is the midpoint of the intensity range. This count was subtracted from the mean value of counts from all images in the dataset and finally squared. This choice of anomaly score generation proved to be reasonable. The results are presented later in the paper. The remainder of this section will provide graphical details of the algorithm in action.

An example of two handwriting samples is shown as images in Figure 1 and Figure 2. These images were produced using the imshow function from the matplotlib python library.

The first step is producing a thresholded image of each sample as. Experiment 1000 resulted in 615 black pixels and 169 white pixels. Experiment 4000 resulted in 705 black pixels and 79 white pixels.

Figure 3: Threshold set to 128 of handwriting sample 1000

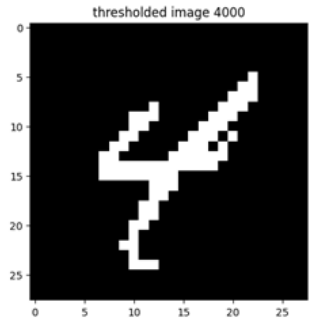

Figure 4: Threshold set to 128 of handwriting sample 4000

Then a mean image is produced to generate the mean count. The count of white pixels for the thresholded mean image is 38. The next step requires subtracting the experiment data from the mean data. The graphical results are shown in the following images.

Now that the subtraction of the experiment image and the mean image has occurred, the resultant value is squared.

3 Results

The probabilities from the data set are in the following output excerpt: Probabilities for Each Class:

Class 3.0: 0.5007 Class 4.0: 0.2503 Class 9.0: 0.1251 Class 6.0: 0.0626 Class 2.0: 0.0313 Class 0.0: 0.0156 Class 5.0: 0.0078 Class 8.0: 0.0038 Class 1.0: 0.0019 Class 7.0: 0.0009

The accuracy of the method was 0.65485, or 65.5

4 Discussion

5 Conclusion