DoubledMNIST: an natural extension of MNIST dataset

Milan Cugurovic
Department of Computer Science
Faculty of Mathematics
University of Belgrade
Belgrade, Serbia, 11000
Email: milan_cugurovic@matf.bg.ac.rs

***Abstract*— The MNIST dataset has become on of the most famous and the most important dataset in machine learning, classification and computer vision task. MNIST was derived from the larger database known as NIST special database 19, which contains handwritten digits, uppercase and lowercase english letters. This document introduce MINST like dataset, but doubled in number of samples and in resolution of each image. After developing an architecture that solve classification task on MNIST dataset, natural learning step is to modify it to solve classification task in similar dataset.**

I. INTRODUCTION

A. NIST Special Database 19

Special Database 19 contains NIST's entire corpus of training materials for handprinted document and character recognition. It publishes Handprinted Sample Forms from 3597 writers, 814,215 character images isolated from their forms. Images are saved in .PNG format. Those segmented characters each occupy 128x128 pixel per raster and are labelled by one of 62 ASCII hexadecimal classes corresponding to "0"- "9", "A"- "Z" and "a"- "z". The characters are given by writer, by class, by caseless class, and by field origin.

B. MNIST dataset

The MNIST database of handwritten digits, published in 1998. by Yan LeCunn et all, is the most popular machine learning database worldwide. It has a training set of 60,000 examples, and a test set of 10,000 examples.
The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

C. EMNIST dataset

The EMNIST dataset is a set of handwritten character digits also derived from the NIST Special Database 19 and converted to a 28x28 pixel image format and dataset structure that directly matches the MNIST dataset. There are six different splits provided in this dataset: ByClass, ByMerge, Balanced, Letters, , Digits and EMNIST MNIST. Images in this dataset follows the same conversion paradigm used to create the MNIST dataset. The result is a set of datasets that constitute a more challenging classification tasks involving letters and digits, and that shares the same image structure and parameters as the original MNIST task, allowing for direct compatibility with all existing classifiers and systems.

II. DOUBLEDMNIST DATASET

The idea for the creation of this dataset is contrary to the idea on which EMNIST was built. As most programmers learn machine learning through MINIST datasets by developing models and algorithms for classification on it, the natural next step, before moving to new datasets, is to modify the eisting architecture to overcome the same problem with some altered data. Which ones double the number, and which are double the larger dimensions.
Cause of that, we derieved this dataset of handwritten english digits. We called it "DoubledMNIST" cause of its dimensions: 140 000 images, resolution 56x56 of each. The conversation paradigm, as in EMNIST, is similar to the technique used in creation of MNIST.

III. METHODOLOGY

A. Conversion Process

The conversion process transforms the 128x128 pixel binary images found in the NIST dataset to 56x56 pixel images with an 8-bit gray-scale resolution that match the characteristics of the digits in the MNIST dataset. For this work we use *ByClass* hiearchy of NIST dataset.
In order to convert the dataset, each digit (we ignore other images) is loaded individually. After that a bounding box is fitted to the character in the image and extracted (using two pixel padding). Then, image is blurred using a Gaussian filter with standard deviation set to 2.

The extracted region of interest is then centered in a square frame with lengths equal to the largest dimension, with the aspect ratio of the extracted region of interest preserved.

Finally, the image is resized to 56x56 pixels using a bi-cubic interpolation algorithm, resulting in a spectrum of intensities which are then scaled to the 8-bit range.

We try different noising and interpolation combination using rudimental Nearest Neighbour algorithm which we train on random subsample of train images and evaluate on random subset of test images. This process is done without any advanced tehnicqe, on pure images as its saved to a dataset. After experiments, we achieved 96.15% precision on pure images, without any preprocessing steps or adwanced methods. Cause of that we choose Gaussian noise with bicubic interpolation as data preparation method.

B. Training and Testing Splits

For this task we also follows the same methodology used in the original MNIST paper in which the original dataset were shuffled and a new random training and testing set were drawn. The training set consist of 120 000 images, every digit corespond to 12 000 of them. Test set is also stratified, and consist of 20 000 images.

IV. CONCLUSIONS

In this paper we provide new dataset which should be the successor to the popular MNIST. This dataset is duplicated in relation to MNIST and switcihg architecture to DoubledMNIST should be real check of knowledge in machine learning tasks. Couse of really good data preparation and dimensions of dataset, with good architecture we expect that precision should be close 1.

APPENDIX

In addition to this dataset, we provide util for unziping it, and load train and test set. Util is saved as .ipynb notebook.