

# Assignment: CNN and MNIST

David Bertoldi – 735213  
email: d.bertoldi@campus.unimib.it

Department of Informatics, Systems and Communication

University of Milano-Bicocca



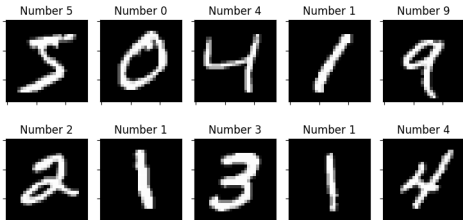
## 1 Inspecting the data

The MNIST dataset contains 70 000 images of handwritten digits (0 to 9) that have been size-normalized and centered in a square grid of pixels. Each image is a  $28 \times 28$  array of floating-point numbers representing grayscale intensities ranging from 0 (black) to 255 (white).

The labels consist of a vector of values, corresponding to the digit classification categories 0 through 9.

The dataset is already divided into training and test sets, respectively with 60 000 and 10 000 samples.

Figure 1 shows an example of the population.



**Figure 1:** The first 10 samples of the train dataset

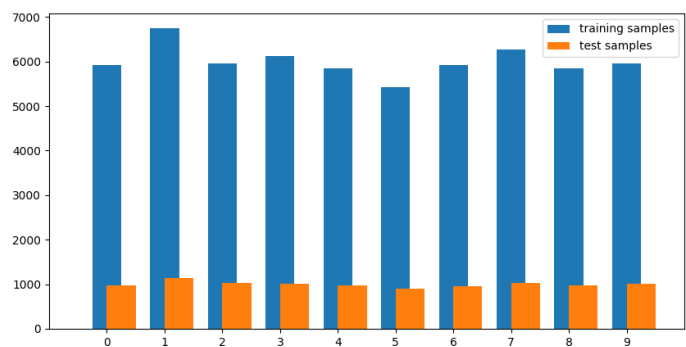
The training population presents a distribution with mean  $\mu = 6\,000$  and standard deviation  $\sigma \simeq 340$  and thus we didn't notice any important unbalance in the data. For this reason we assumed the data followed a distribution  $X \sim U(\mu, \sigma)$  and no data augmentation on less populated classes was taken into account. Figure 2 shows the data distribution for both training and test datasets.

## 2 Preparing the data

Before training a FFNN using this images, encoded in  $28 \times 28$  matrices with values from 0 to 255, we flattened them in arrays  $1 \times 784$  and rescaled each value in the continuous interval  $[0, 1]$ . This encoding will be used in every section of this work: a flat array better suits the input layer of a FFNN and small values increases the efficiency in the calculations.

### 2.1 Data split

As noted in section 1, the dataset is divided into training and test samples. A validation subset is missing and thus



**Figure 2:** Histogram of the frequency of samples in the dataset

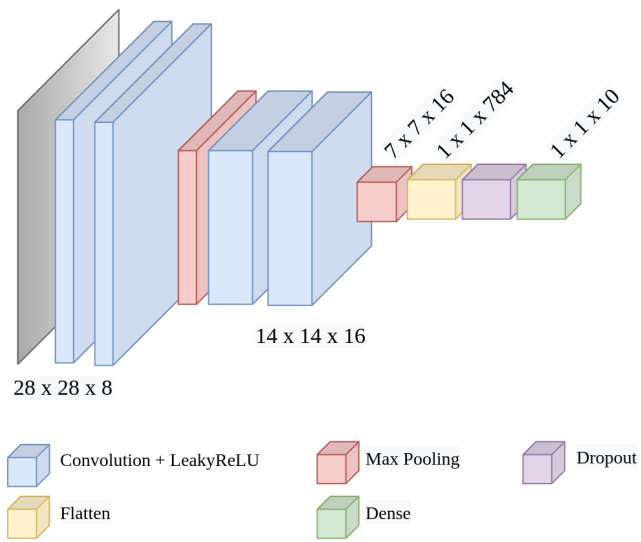
is retrieved from the training set: 15% of the images are randomly used for validation instead of training (along with their labels) for a total of 9 000 samples.

About labels, we encoded them in one-hot vectors so that the 1s are set in the index representing the numerical class.

## 3 Building the network and training

The aim of this section is to describe a CNN with less than 10 000 parameters that is able to classify with high level of accuracy the numbers from the dataset with or without any regularization technique.

### 3.1 The network



**Figure 3:** Histogram of the frequency of samples in the dataset