Ablation study: Achieving 99.47% accuracy in MNIST using MLPs

Miralles Ferrer, Francisco

1 Introduction

This work presents an ablation study performed on the MNIST test set, with the aim of analyzing, both individually and collectively, the impact of different proposed components on the performance of a baseline model. The goal is to achieve a minimum accuracy of 99.4% on the MNIST test set.

MNIST is a widely used database in machine learning, composed of images of handwritten digits. Each image is 28x28 pixels in size and is labeled in one of 10 classes, corresponding to the digits 0 through 9. The dataset is divided into 60,000 training images and 10,000 test images. Examples of MNIST images are shown in **Figure 1**.

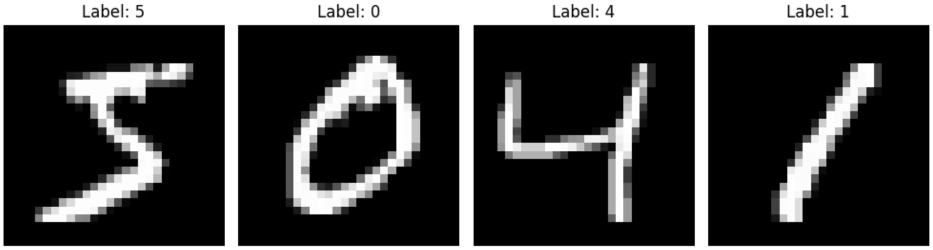

Figure 1: MNIST Example

As an additional restriction of the present work, the use of convolutional layers is prohibited, which forces us to improve the performance of the baseline model through alternative techniques, such as redesigning the architecture based on multilayer perceptrons and employing data preprocessing and augmentation strategies.

2 Baseline

As a starting point, a baseline model based on a multilayer perceptron (MLP) is used, composed of four linear layers with ReLU activation functions. The first layer projects the 784 input features, corresponding to the vectorized input image, into a 1024-dimensional space. This size is maintained in the two intermediate hidden layers, while the final layer reduces the dimensionality to 10 outputs, one for each class of the classification problem.

The baseline pipeline does not incorporate any data augmentation techniques. During training, it is used a stochastic gradient descent (SGD) optimizer, with a learning rate of 0.01, a weight decay of 1×10^{-6}, a momentum coefficient of 0.9 and cross entropy loss.

3 Added components

To improve the baseline model's performance, modifications have been made to two fundamental aspects. First, an alternative architecture has been proposed that incorporates additional training regularization and stabilization techniques. Second, preprocessing and data augmentation strategies have been added to enhance the model's generalizability.

The impact of each of these modifications is analyzed individually and jointly through an ablation study, which allows for the evaluation of their specific contribution to the final performance on the set of tests.

3.1 Proposed model

An architecture has been proposed for the MLP model in which batch normalization techniques, dropout, and residual connections have been added.

The architecture is shown in **Figure 2**. First, an initial block is applied, consisting of a linear layer that converts the 784 input features to 1024, a batch normalization layer, a ReLU activation function, and moderate dropout. Next, two residual blocks are applied, each containing a ReLU activation function after adding the residual connection. Finally, a linear layer is used to convert the 1024 features to the 10 output features.

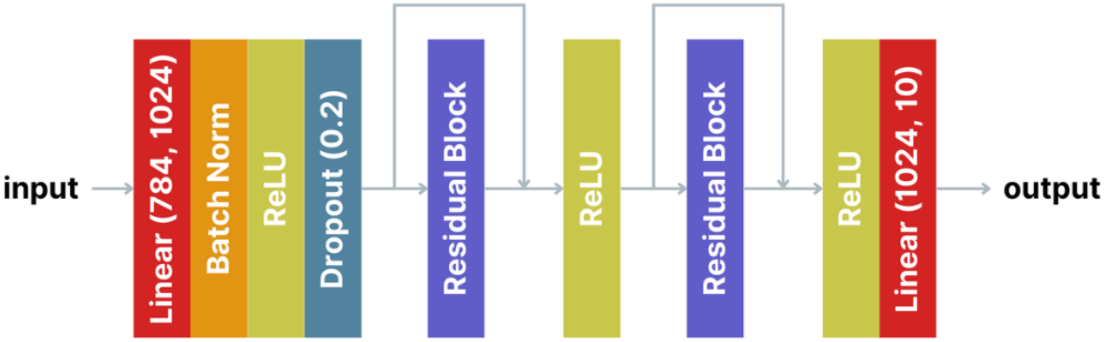

Figure 2: Architecture of the model

Regarding the developed residual block, its architecture is as shown in **Figure 3**. It consists of a first linear layer that maintains the size of 1024 features. Next, it applies batch normalization and a ReLU activation function. Finally, it uses another linear layer and another batch normalization.

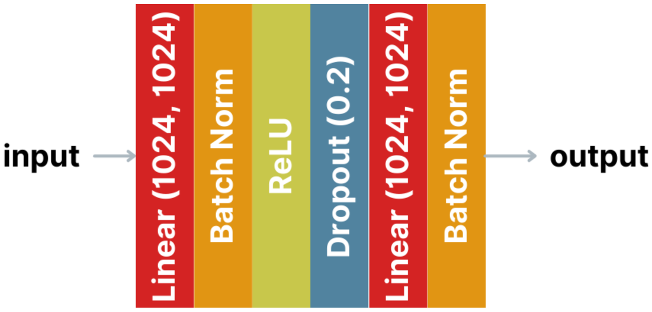

Figure 3: Architecture of the residual block

3.2 Proposed data augmentation

A data augmentation pipeline is proposed, designed to improve the model's generalizability across the MNIST dataset. The preprocessing applied to the training and test sets is explicitly differentiated:

- **Training:** First, smooth random rotations are applied over a range of ±10°, with the aim of introducing some slight variation without altering the semantics of the digits (for example, steeper rotations could induce confusion between classes such as 6 and 9). Then, affine transformations are employed, including translations of up to 10% of the image and scaling over the range $[0.9, 1.1]$, which increases the robustness of the model against changes in the position and size of the digits.

  Subsequently, elastic deformations are applied, a common technique in MNIST that simulates natural variations in handwriting. After these transformations, the image is converted to a tensor, normalizing its values to the range $[0, 1]$. Finally, Gaussian noise with a mean of zero and a standard deviation of 0.05 is added, and normalization is performed using the mean and standard deviation of the MNIST dataset, which helps to stabilize the training process.

- **Test:** The images are converted to tensor and normalized using the same mean and standard deviation used during training, thus ensuring a consistent evaluation of the model.

4 Ablation study

It is presented an ablation study that analyzes the impact on the accuracy obtained on the test set using the proposed components. Each of these components has been evaluated in 100 epochs.

Regarding the nomenclature used, *Model A* corresponds to the baseline model, while *Model B* refers to the proposed model described in the previous section. Similarly, *Preprocessing A* denotes the preprocessing used in the baseline, and *Preprocessing B* the new proposed pipeline, which incorporates data augmentation techniques. The combination of these two factors results in four experimental configurations, the results of which are summarized in **Table 4**.

Configuration	Model	Preprocessing	Accuracy (%)
Baseline	A	A	98.31
Model change	B	A	98.83
Preprocessing change	A	B	99.40
Model + Preprocessing	**B**	**B**	**99.47**

Table 1: Ablation study results.

5 Conclusions

The results obtained in the ablation study allow us to draw several relevant conclusions. The baseline model achieves an accuracy of 98.31% across the test set. Replacing the original architecture with the proposed model produces a moderate improvement, raising the accuracy to 98.83%. Furthermore, applying the proposed preprocessing in conjunction with the baseline model provides a substantial performance increase, achieving an accuracy of 99.40%, thus meeting the initially set objective.

Finally, the combination of the proposed model and the new preprocessing leads to the best observed result, with an accuracy of 99.47% in the test set. These results demonstrate that, while improvements in the architecture contribute to system performance, preprocessing and data augmentation techniques play a fundamental role in the model's generalizability, allowing it to exceed the established accuracy threshold.