Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Dissecting the Neural Code: Neuron-by-Neuron Analysis of a Digit Recognition Network

Vitor H. M. Mourão[1], João Batista Florindo[2]
UNICAMP, Campinas, SP

Abstract. This study analyzes the role of individual neurons in neural network performance, focusing on the concept of "rotten" neurons whose removal affects network accuracy. By examining two neural network configurations across the MNIST and SVHN datasets, we demonstrate the diverse impact of neurons, from beneficial to detrimental. Our results reveal that neural network efficiency can be improved by addressing the influence of specific neurons. This research highlights the potential for neuron-level analysis and pruning to enhance neural network optimization.

Keywords. neuron-level analysis, neural network pruning, neuron importance

1 Introduction

Neural networks are powerful tools for pattern recognition and prediction, yet the significance of individual neurons within these networks is not fully understood. In this study, we perform a comprehensive analysis by assessing the impact of the removal of individual neurons on the performance of the network. To conduct this analysis, we utilized well-known datasets, namely the MNIST [1] and the Street View House Numbers (SVHN) [2], and employed a Multilayer Perceptron (MLP) Classifier using the "sklearn" package [3]. This approach allowed us to systematically investigate how neurons contribute to the overall functionality and performance of the network. Although the concept of pruning neural networks to enhance performance and efficiency is already a popular discussion in the field [4, 5], the in-depth analysis of the effects of individual neurons, particularly in terms of their unique contributions and potential redundancy, is not thoroughly explored in existing literature. This gap highlights the novelty of our research, which, through a series of experiments involving neural networks with varying configurations, elucidates the critical role individual neurons play in complex pattern recognition tasks. Our findings offer insights into the architectural nuances of neural networks and their operational dynamics, providing a more granular understanding of neural network architecture and functionality.

2 Methodology

In this study, we undertook a detailed analysis to quantify the impact of individual neurons on the overall performance of neural networks. The first neural network configuration, referred to as NN_392, consists of a single hidden layer with 392 neurons, which is half the size of the input layer for the MNIST dataset. The second configuration, NN_392_196, includes two hidden layers, with the first layer containing 392 neurons and the second layer 196 neurons. These configurations were selected to explore the effects of network complexity and depth on neuron significance.

[1] v137856@dac.unicamp.br
[2] jbflorindo@ime.unicamp.br

We conducted experiments using two well-known datasets: the MNIST and the SVHN. For the MNIST dataset, we utilized a total of 70,000 instances, splitting them into 63,000 for training and 7,000 for testing, adhering to a 90/10% holdout. Similarly, for the SVHN dataset, we worked with a total of 99,289 instances, allocating 73,257 for training and 26,032 for testing, reflecting a 74/26% holdout. Each neural network configuration was trained separately on these datasets, utilizing the scikit-learn package's MLPClassifier. After training, we "removed" individual neurons from each configuration by setting their weights and biases to zero and then proceed to analyze the accuracy change on the network without that neuron. This process was repeated for all hidden neurons and allowed us to measure the impact of each neuron's removal on the network's accuracy on a fixed test set for both the MNIST and SVHN datasets.

Our analysis extended beyond merely assessing the influence of individual neurons. We also visualized the weight distribution of the most "rotten" or least contributing neurons, compared the influence of neurons across different layers, and examined the impact of each synapse of the identified "rotten" neurons. Through this comprehensive approach, we aimed to shed light on the intricate dynamics of neural network functionality and the critical role of individual neurons in complex pattern recognition tasks.

3 Results

3.1 Model NN_392

The visualizations presented in Figures 1 and 2 effectively illustrate the differential impact of individual neurons on the model's predictive accuracy for the MNIST and SVHN datasets, respectively. Each neuron's influence is plotted in ascending order, revealing the diverse spectrum of contributions towards the network's performance.

For the MNIST dataset, the network demonstrated a high training accuracy of 99.78%. With all neurons active, the test accuracy was recorded at 98.06%, which slightly improved to 98.10% upon the removal of the most "rotten neuron". This subtle increase hints at the presence of specific neurons that detract from overall performance. Out of the 392 neurons, analysis revealed that 255 neurons (65.05%) had a positive impact, 106 neurons (27.04%) had no impact, and 31 neurons (7.91%) had a negative impact on the network's accuracy.

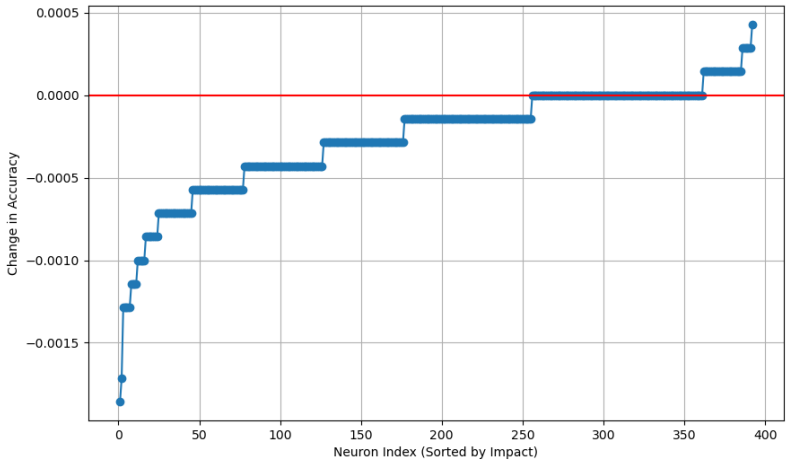

Figure 1: Impact of Neuron Removal on Accuracy on the MNIST dataset. The horizontal red line denotes no change in accuracy. Source: Author.

In contrast, the SVHN dataset, with its larger and more complex color images, presented different challenges, as reflected by the lower training and test accuracies of 65.07% and 59.83%, respectively. In this network configuration, 392 neurons were analyzed, with 10 neurons (2.55%) showing a positive impact, 382 neurons (97.45%) showing no change, and notably, 0 neurons (0.00%) showing a negative impact on accuracy. This uniformity in the neuron impact distribution suggests that for more intricate datasets like SVHN, individual neurons may not exhibit detrimental effects on performance as noticeably due to the network's limited capacity to encapsulate the complexities of the data fully.

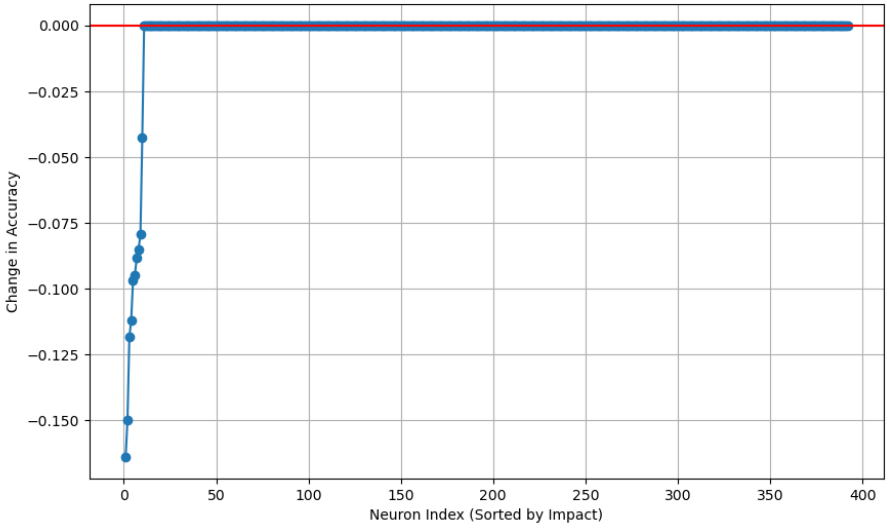

Figure 2: Impact of neuron removal on accuracy on the SVHN dataset. The horizontal red line denotes no change in accuracy. Source: Author.

In the MNIST dataset, a detailed examination was conducted on Neuron 137, which was identified as having the most substantial negative impact on network performance. Termed a "rotten" neuron, it was uniquely characterized by its consistent contribution to degrading the model's predictive accuracy. A visualization of Neuron 137's weights, formatted as a 28x28 image (Figure 3), revealed a pattern analogous to a circled cloud of mixed pixels. This pattern lacks the coherent features typically associated with effective predictors and is indicative of its negative influence on the network's classification ability. In contrast, such a distinctive "rotten neuron" was not observed in the SVHN dataset, where no single neuron demonstrated a clear and singular adverse effect on the network's accuracy.

Additionally, we identified four specific instances where Neuron 137 involvement led to incorrect classifications (Figure 4). These instances were particularly challenging, even for human interpretation, supporting the notion of the neuron's detrimental effect.

Subsequent experiments involved deactivating individual synapses (weights) of the "rotten neuron" to assess their impact (Figure 5). This granular approach revealed that certain synapses significantly contributed to the poor performance of the neuron, although not to the same extent as the whole neuron.

DOI: 10.5540/03.2025.011.01.0382

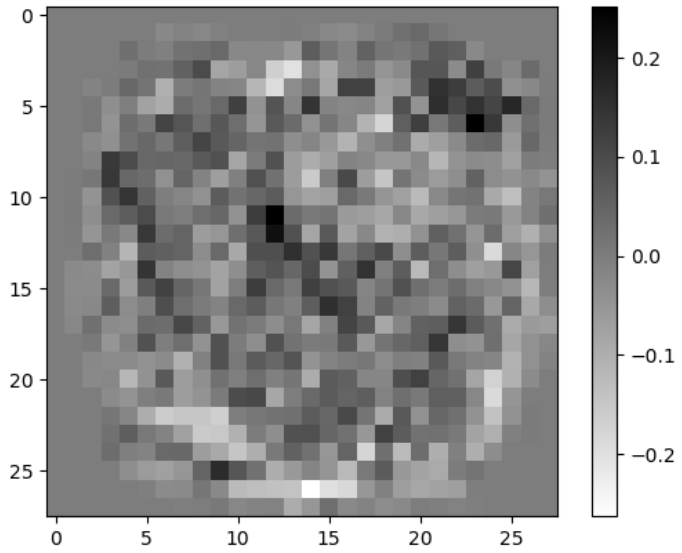

Figure 3: Weights of Rotten Neuron 137. Source: Author.

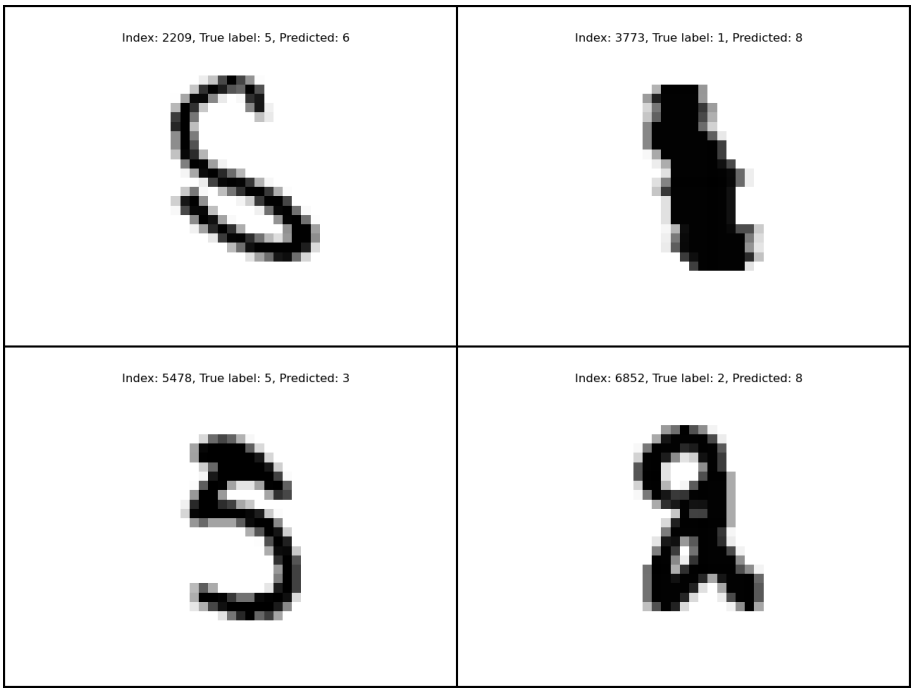

Figure 4: Cases where 'Neuron 137' makes the model classify incorrectly. Source: Author.

DOI: 10.5540/03.2025.011.01.0382

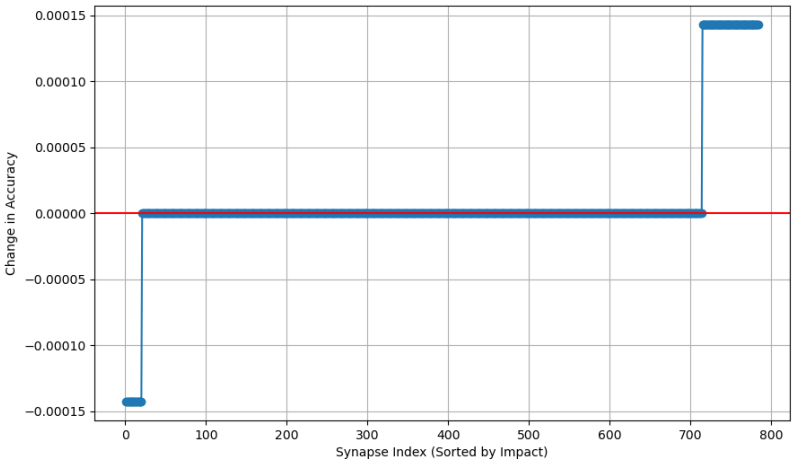

Figure 5: Impact of Synapse Removal on the NN_392 model (MNIST). Source: Author.

3.2 Model NN_392_196

In the investigation of the NN_392_196 configuration, the complexity of having an additional layer was examined. The results, depicted in Figures 6 and 7, outline the differential impact of neuron removal on the accuracy of the model for both the MNIST and SVHN datasets.

For the MNIST dataset, the model achieved near-perfect training accuracy (99.95%) and high test accuracy (98.21%). Upon the sequential deactivation of neurons, the first layer presented a mixed influence on the model's accuracy, with 39.29% of neurons leading to a decrease in accuracy when removed, and 19.90% showing a positive impact when removed. The second layer, however, had a more pronounced positive impact upon neuron removal, with 36.73% of neurons showing an increase in test accuracy when deactivated and 39.80% showing a decrease. The most significant positive change in accuracy from a single neuron removal was greater in the second layer (0.0011) compared to the first (0.0007), as illustrated in Figure 6.

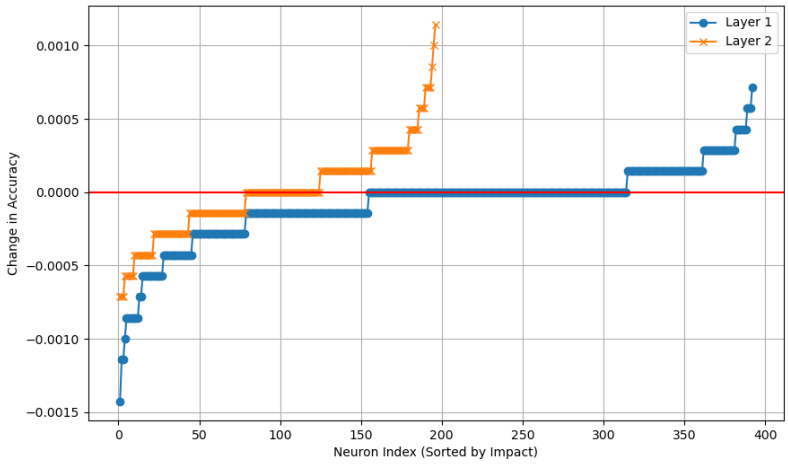

Figure 6: Impact of Neuron Removal, on both layers, on Accuracy on the MNIST dataset. The horizontal red line denotes no change in accuracy. Source: Author.

Conversely, the SVHN dataset, which poses more complex image recognition challenges, exhibited distinct neuron impact patterns when analyzed with the NN_392_196 configuration. Only 5.61% of the neurons in the first layer decreased the model's accuracy upon their removal. Notably, no neurons in the first layer enhanced performance when deactivated. In the second layer, however, 10.71% of neurons showed a positive impact on accuracy when removed, while a substantial 77.55% led to a decrease in accuracy. Furthermore, the most considerable positive change in accuracy was observed in a neuron from the second layer (0.0022), which is notably double the maximum change observed in the NN_392_196 MNIST model. However, this change might be misinterpreted due to the different scales used in Figures 6 (MNIST) and 7 (SVHN). The addition of a second layer also notably improved the accuracy of the SVHN model, resulting in a training accuracy of 87.94% and a test accuracy of 80.31%.

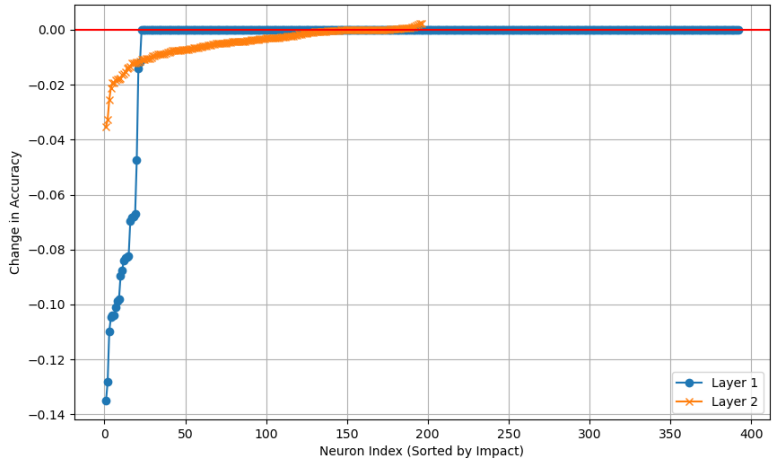

Figure 7: Impact of Neuron Removal, on both layers, on Accuracy on the SVHN dataset. The horizontal red line denotes no change in accuracy. Source: Author.

These observations suggest that in more complex network architectures, like NN_392_196, the role of neurons in each layer may be more distinct, with the second layer possibly playing a more critical role in refining the network's output. The "rotten" neurons, particularly within the second layer, have a more substantial impact on the model's predictive accuracy for both datasets, underscoring the importance of considering layer-specific dynamics in neural network analysis.

The analysis of the NN_392_196 configuration across both the MNIST and SVHN datasets has elucidated the distinct roles that neurons play in each layer of a neural network. The second layer's neurons, in particular, have been highlighted as pivotal in refining the network's predictive accuracy, especially for more complex datasets. These insights underscore the delicate balance required in network architecture design, where the depth and distribution of neurons must be tailored to the complexity of the task at hand.

4 Conclusions

Our investigation into the role of individual neurons within neural networks has uncovered a complex landscape of impacts across various configurations and datasets. We have identified neurons that are pivotal to network performance, alongside those that appear neutral or even detrimental—coined as "rotten" neurons.

DOI: 10.5540/03.2025.011.01.0382

The concept of "rotten" neurons, primarily explored within the context of single-layer and two-layer perceptrons in this study, opens the door for extensive research into more sophisticated architectures. The potential existence of analogous "rotten" entities, such as filters in convolutional neural networks (CNNs), cells in recurrent neural networks (RNNs), and heads in attention mechanisms, suggests a universal paradigm that could influence a wide array of neural network applications. Investigating these could yield crucial insights into optimizing feature extraction, sequence processing, and model interpretability, respectively.

Looking forward, the methodology applied in this research offers a framework for further explorations into neural network inefficiencies. Future studies could extend our approach to various neural network architectures, employing advanced techniques to systematically identify and mitigate the effects of "rotten" computational units. Such endeavors could not only enhance model performance but also contribute to the creation of AI systems that are both more robust and resource-efficient.

References

[1] Yann LeCun, Corinna Cortes, and Christopher J Burges. **The MNIST database of handwritten digits**. Online. Accessed on March 8, 2024, `http://yann.lecun.com/exdb/mnist/`. 1998.

[2] Yuval Netzer and Tao Wang. "Reading digits in natural images with unsupervised feature learning". In: **NIPS workshop on deep learning and unsupervised feature learning.** (2011). Vol. 2011, No. 5.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: **Journal of Machine Learning Research** (2011). Vol 12. pp. 2825-2830.

[4] Huan Wang, Can Qin, Yue Bai, and Yun Fu. **Why is the State of Neural Network Pruning so Confusing? On the Fairness, Comparison Setup, and Trainability in Network Pruning**. 2023. DOI: `https://doi.org/10.48550/arXiv.2301.05219`.

[5] Xin Yu, Thiago Serra, S. Ramalingam, and Shandian Zhe. **The Combinatorial Brain Surgeon: Pruning Weights That Cancel One Another in Neural Networks**. 2022. DOI: `10.48550/arXiv.2203.04466`.

